

**Transcriptional Analyses of the
Infoture Natural Language Corpus**

Kimberly Coulter, Jill Gilkerson, & Jeffrey A. Richards

Infoture, Inc., Boulder, CO

ITR-06-1

October 2007

LENA™ Hardware Model: LR-0120 Software Version: V2.3.0

Copyright © 2007, Infoture, Inc. All Rights Reserved

The LENA™ System

The LENA language environment analysis system is a language monitoring and feedback system designed to provide information about the language environment of infants and toddlers to parents, clinicians, and researchers. The LENA System includes the LENA digital language processor (DLP) that children ages 2 through 36 months wear in the pocket of custom-made clothing. It records everything the child says and hears over a continuous 16-hour day. The audio data is transferred to a computer and analyzed by the LENA language environment analysis software. Parents can access automatically generated feedback reports to view objective information about their child's language environment. The Adult Word Count (AWC) report provides estimates of the total number of adult words the child hears, and the Conversational Turns (CT) report provides estimates of the total number of conversational interactions the child engages in with an adult. These reports permit AWC and CT estimates to be viewed as hourly, daily, or monthly totals. Daily AWC and CT percentile ranking estimates based on a normative database are reported in the LENA software.

The LENA System is intended: 1) to provide a measurement tool to help researchers gain insight into the natural language environment of children; 2) to aid professionals in the early detection of language delay; 3) to support home intervention programs directed at improving the language environment of language-delayed or disadvantaged children; and 4) to educate and provide feedback to parents regarding how much they talk to and interact with their children in order to aid them in maintaining and improving their children's language environments.

Abstract

Quantifying the accuracy and reliability of the LENA™ language environment analysis software V2.3.0 is dependent on accurate and reliable transcription of the audio files by professional transcribers. Here, we disclose the analytical procedure designed and implemented by the Infoture transcription team to identify and code speakers. Very good inter-rater reliability was achieved between the criterion rater and seven secondary raters. Inter-rater reliability was assessed through segment classification agreement. Significant accuracy in the transcriptions provided by professional raters was critical to train the processing models used to automatically identify segments in the audio file and to subsequently test the accuracy and reliability of the segmentation process.

Keywords

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

1. Introduction

Audio data collected by the LENA™ Digital Language Processor (DLP) is transferred to a computer and analyzed by the LENA language environment analysis software V2.3.0. The LENA software is a compilation of statistical models and algorithms that were trained to extract and segment features of the audio data. The software models were trained using professional transcriptions. Thus, the reliability of the professional transcripts was essential to model development and refinement. In this technical paper, we describe the transcription process and discuss the level of inter-rater reliability that was achieved.

2. Description of the Database

In 2005, scientists at Infoture started a large-scale study to assess the language environments of 2- to 36-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 natural home environments, known officially as the Infoture Natural Language Corpus. The current database consists of nearly 40,000 hours of audio from over 300 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 Infoture Natural Language Corpus to develop and test the speech processing algorithms that contribute to the LENA software V2.3.0. In addition, the transcription data collected contributed to the development of normative estimates and to other research regarding the natural language environment of children.

2.1 Methodology

The professional transcription team at Infoture 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.

The primary purpose of the Infoture transcriptions was to optimize the LENA software V2.3.0 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 ITR-05-1 for information about the reliability of the LENA System.

2.2 Segment Identification

Overview

For the purposes of segment identification, audio data recorded by the LENA DLP were segmented into first tier components that included adult male (near and distant), adult female (near and distant), key child, noise, overlapping speech, other child, and media noise. The transcribers segmented instances of ≥ 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. Transcribers also identified conversational

turns (interactions between key child and adult) through back and forth alternation between the key child and an adult speaker. A conversational turn occurred if there was <5 sec of pause between speakers. Please see Technical Report ITR-04-1 for further information on identification of conversations and measuring conversational turns.

General Identification Codes

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

Identification Code Subcategories

Many of these 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 Infoture transcription team assigned 31 general codes for audio file analyses. Each of the 31 unique identification codes were subject to further classification based on sound wave amplitude. Low amplitude regions were identified as far-field signals, indicating the speaker was distant from the key child. Segments containing far-field signals were not included in the word count

processing since the meaningful contribution of these segments to the child's language environment was considered to be minimal. Once the adult near-field segments were identified and segmented, Infoture 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 ITR-05-1.

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 contained sounds that were instinctive emotional reactions to the environment (e.g. crying, screaming, laughing). Vegetative sounds were those sounds resulting from respiration (e.g. breathing) or digestion (e.g. burping) (Oller, 2000).

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 the periods of time during which the child was in a car or at daycare.

2.3 File selection

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

2.3.1 Training Set

Large-scale training sets derived from the Infoture Natural Language Corpus were used to test and 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 each child's age and gender. Approximately five male and five female children ranging in age from 1-42 months were selected per each month of age. The final sample included 157 male and 152 female children.

2.3.2 Test set

The primary purpose of the transcription test set generation was to assess the accuracy and reliability of the LENA software V2.3.0 using a representative sub-sample of the Infoture 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 (2-36 months), gender, SES, and independent standard assessments of the recorded child's language ability (e.g. *Preschool Language Scale-4* (Zimmerman, Lee, Steiner, & Pond, 2002)). Two children were selected per each month age group (e.g. 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 software engineers from the Infoture research and development team 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 speakers. For later analyses, these six ten-minute segments were concatenated to form one hour of transcription per file.

3. Inter-rater reliability

Inter-rater reliability was assessed using audio files obtained from one 11-month-old female and one 25-month-old male. Inter-rater reliability analyses compared a criterion rater with seven secondary raters on these files. These analyses were based on two recording sessions with 1731 unique criterion-identified segments that further included 429 criterion-identified child speech segments. Three separate analyses were conducted to assess inter-rater agreement with respect to: adult versus non-adult speech; child versus non-child speech; and for the key child, speech versus non-speech.

3.1 Segment Boundaries

To assess inter-rater reliability, the criterion rater's segments established the official segments of interest. Secondary raters' segments were compared to segments established by the criterion rater using time codes to identify the start and stop points of each segment. Because of potentially differing segment boundaries, it was sometimes the case that one criterion-based segment could span several secondary rater segments. In these data the criterion-based segments spanned a maximum of four secondary rater segments. However, on average nearly 70% of criterion-based segments matched secondary rater segments exactly, and an additional 28% spanned only two segments (Table 1). Thus, the data reveal a high degree of segmentation agreement.

Table 1: The Number of Secondary Rater Segments Spanned by One Criterion Rater Segment

Rater	Number of Segments Spanned				Total Segments
	One	Two	Three	Four	
1	1155	524	49	3	1731
2	1257	436	37	1	1731
3	1173	491	65	2	1731
4	1234	469	28	0	1731
5	1070	574	87	0	1731
6	1313	393	25	0	1731
7	1224	471	36	0	1731
Mean	1204	480	47	1	1731

3.2 Classification Agreements

Classification agreements (hits) were defined as occurring when any portion of secondary raters' segments overlapped with the criterion rater's segments. Classification agreement was computed separately for identification of adult versus non-adult and key child versus non-key child segments. Within criterion-identified key child 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 overall (unadjusted) agreement and as the chance-corrected coefficient kappa, κ (Cohen, 1960). Results of the agreement rating analyses are shown in Table 2.

Table 2: Overall Agreement with associated kappa (κ) values for Adult versus Non-Adult Speakers; Key Child versus Non-Key Child speakers; and Key Child Speech versus Key Child Non-Speech

Rater	Adult vs. Non-Adult Speakers	Key Child vs. Non-Key Child Speakers	Key Child Speech vs. Key Child Non-Speech
	Overall Percent (κ)	Overall Percent (κ)	Overall Percent (κ)
1	.96 (.79)	.94 (.84)	.92 (.82)
2	.96 (.75)	.91 (.75)	.89 (.77)
3	.96 (.77)	.90 (.72)	.86 (.70)
4	.96 (.80)	.93 (.81)	.94 (.86)
5	.97 (.80)	.94 (.83)	.89 (.76)
6	.96 (.74)	.91 (.75)	.83 (.66)
7	.96 (.79)	.92 (.78)	.85 (.69)
N	1731	1731	429

As Table 3 demonstrates, secondary raters' identification of adult versus non-adult segments overlapped the criterion rater's segments approximately 96% of the time. Similarly, secondary raters' segments of key child versus non-key child speakers overlapped the criterion rater's segments on average 92% of the time. When the raters identified segments containing key child speech versus key child non-speech, they were in agreement approximately 89% of the time. Thus, the level of agreement among raters across the categories of interest demonstrated a degree of accuracy more than sufficient for these transcriptions to be used for model training and testing.

4. Conclusion

Professional transcriptions of audio data were used as a basis for statistical model building and refinement for the LENA software V2.3.0. Infoture's professional transcription team developed a detailed and highly analytical system for speaker identification and segmentation. A training set was

transcribed professionally and was used to train statistical models developed by the Infoture engineering team for algorithmic analysis of audio data; a separate sub-sample of the Infoture Natural Language Corpus was used as a transcription test set to assess the accuracy and reliability of the LENA software V2.3.0. The inter-rater reliability of the transcription team was very high, as apparent from segmentation boundary data and classification agreements. This level of reliability was critical for fine-tuning the LENA software.

References

- Cohen, J.A. (1960) A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, 20, 37-46.
- Oller, D.K. (2000) *The Emergence of the Speech Capacity*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- U.S. Census Bureau (2005) *Current Population Survey*. Washington, DC: U.S. Government Printing Office.
- Zimmerman, I.L., Steiner, V.G., Pond, R.E. *Preschool Language Scale, Fourth Edition*. San Antonio: The Psychological Corporation, 2002.