LENA®

A Guide to Understanding the Design and Purpose of the LENA® System

Jill Gilkerson & Jeffrey A. Richards
LENA Foundation, Boulder, CO

LTR-12

July 2020

Increasingly over the past decade, validation studies have been published that evaluate the accuracy of the LENA System™. This document provides a framework for critically reviewing such studies and a reference to inform future validation efforts. It is crucial that researchers and practitioners assessing LENA be aware of specific complexities directly related to its core design and goals for the report metrics.

CORE FEATURES OF THE LENA SYSTEM

The LENA System was designed foremost to provide insight into patterns of child vocal behavior and development by equipping caregivers with information about the frequency with which they interact with children in their care. That is, the primary goal was to generate simple, high-level feedback on a child's natural language environment to promote adult behavior change. To achieve this goal, it was necessary first to identify and distinguish child from adult vocalizing. **Simply put, the system was optimized to identify with high accuracy vocalizations from: 1) the child wearing the recorder, and 2) nearby adults, and to eliminate everything else.**[1] A second goal was to generate reliable estimates of the frequency of Adult Words (Adult Word Count), adult-child alternations (Conversational Turns Count) and Child Vocalizations (Child Vocalization Count). Finally, given that LENA offers a new window into the daylong language experience of very young children, an additional goal was to establish a meaningful context by which to interpret these metrics via comparison to a normative sample.

LENA technology does not attempt to recognize or understand the meanings of words. Rather, once adult speech is identified, the algorithm estimates the number of words spoken based on specific information in the speech signal, such as syllable count, consonant distribution, and segment duration. LENA algorithms utilize transcription-based sound models to map or segment each moment of the audio recording stream onto eight unique sound categories, a process referred to here as "segmentation labeling." Four primary segmentation labels contribute directly to LENA measures: Key Child (the child wearing the recorder), Adult Female, Adult Male, and TV/Electronic media. The remaining four secondary segmentation labels are not utilized for core reports: Other Child, Overlap, Noise, and Silence. (See Appendix A for operational definitions of the eight segmentation labels and the core LENA measures.) Note that during the algorithm design phase of LENA's development,

[1] Given guidance from the AAP that adults should minimize television exposure for young children, it was also necessary to provide information about exposure to television in the TV/Electronic Sounds report.

accurate labeling of the *primary* segmentation labels was prioritized, and focus on the unutilized *secondary* segmentation labels was correspondingly minimized. For example, LENA's developers were less concerned about how well Other Child was differentiated from Noise, since neither would contribute to the core feedback reports.

ORIGINAL VALIDATION OF LENA PERFORMANCE

LENA's segmentation accuracy was initially evaluated on 70 hours of human transcription coding completed on an age- and gender-balanced sample of typically developing children 2 months to 36 months of age (the original age range of interest) living in monolingual North American English-speaking households. Six 10-minute sections were selected and coded from each recording, and results were reported in Xu et al., 2008. (See also LENA Technical Report LTR-06-2: *Transcriptional Analyses of the LENA Natural Language Corpus*[2] for a description of the coding procedures.) Later, child modeling was extended to 48 months, and an additional three 10-minute sections from each of 24 recordings (12 hours) from children 37 months to 48 months were selected, transcribed, and added to the performance evaluation set, for a total of 82 hours from 94 children. See Appendix B for the full sensitivity and precision confusion matrices, including all eight segmentation labels.

Since the LENA technology's release in 2008, many independent researchers have recognized its unique potential as a tool to study naturalistic behaviors and to identify children for intervention services. In some cases, such applications have utilized the system in contexts outside its original validation parameters (e.g., beyond the established child age range or home language), and evaluators correctly have attempted to validate its accuracy in these specific contexts. Unfortunately, these efforts sometimes have incorporated and further propagated erroneous assumptions regarding the system design, which we attempt to clarify here. We strongly encourage readers of published LENA studies, reviewers of newly submitted articles, and researchers designing novel LENA validation studies to consider the information provided herein and to reach out to the LENA team with any and all questions.

[2] https://www.lena.org/wp-content/uploads/2016/07/LTR-06-2_Transcription.pdf

RECORDING LOGISTICS: CLOTHING, THE RECORDER, AND RECORDING ENVIRONMENT & DURATION

Segmentation category models were trained on audio from daylong recordings made in a child's natural language environment (generally the home) using the LENA recorder specifically positioned in LENA clothing. LENA clothing material was selected for its low friction properties, and optimal recorder pocket placement was determined empirically. It follows that LENA's segmentation category models may be sensitive to recorder-related acoustic features specific to recording conditions — for one, Key Child labeling accuracy depends in part on the recorder's proximity and unimpeded access to the child's mouth to capture "live" vocalizations. Consequently, validation of LENA labeling and metrics should *never* encompass playing pre-recorded audio into a LENA recorder; electronically reproduced vocalizing is intended to be labeled as TV/Electronic Sounds. Similarly, although the acoustic modeling is derived from recordings made in a natural speech environment (which could include time spent outdoors), segmentation labeling should not be expected to be as accurate when, for example, a child is outside on a windy day with the recorder under a coat.

In general, validation of recording environments should encompass all the usual ambient household noises that may arise over the course of a typical day. Sampling from across a daylong recording helps to ensure that relatively brief, anomalous environmental conditions (e.g., a dog barking or a lawn mower running outside the window) do not have an inordinate impact on a performance evaluation. LENA technology can provide a "long exposure" rendering of a child's language environment that provides more stable estimates, and thus validation samples should be drawn from full, daylong recordings, not, e.g., be restricted to selections from short recordings made in controlled environments.

CHILD AGE & PUBERTY

Identification of Key Child vocalizations utilizes age-specific modeling derived from the vocal output of children 2 months to 48 months old. When a person beyond this age range wears the LENA recorder, LENA algorithms will reference the model for a 48-month-old. The degree to which this would be problematic of course varies with each situation. Whereas the acoustic signature of a 50- or 60-month-old child may be relatively similar to that of a four-year-old, the vocal tract of a person past puberty is quite different acoustically. Validation work should never incorporate audio samples from post-pubescents wearing the LENA recorder, or at a minimum their recording data should not be combined with validation results from children under 48 months of age.

SELECTING AUDIO SEGMENTS FOR VALIDATION

The standard goal of audio segment selection to evaluate LENA performance should be to collect an unbiased sample representative of a child's entire recording day. If generating several hours of coding from within a single day is feasible, then selection may be random from continuous 10-minute sections (avoiding sleep time; see below). When only an hour or less of a given day is coded, researchers should attempt to sample at varying levels of speech activity to obtain a realistic distribution of typical events and errors. (See Recommended Guidelines for Validating LENA Word and Turn Counts, available from info@LENA.org, for more information on selecting low-, mid-, and high-activity regions). Recording regions that LENA indicates as having high activity are important to include, but sampling high-activity periods exclusively may also generate high error rates that are unrepresentative of the rest of the day. For example, an inflation of turns counts could occur if LENA were to misidentify an adult using parentese as a child, but unless that speech itself is typical of what the adult produces all day, the resulting error rate for turns counts will be artificially high.

As well, validation audio sampling should not be based solely on recordings from a fixed time point each day, which would be more likely to include similar activities and thus bias error rates. Similarly, it is advisable to avoid sampling audio from the beginning and end of a recording exclusively due to a range of confounding factors that can potentially affect behavior and bias error rates, such as novelty effects during the first hour and fatigue at the end of the day. The goal should always be to sample from across a range of activities and timeframes that more fully represents the language environment overall.

SILENCE

There are two important issues related to Silence detection to consider in validation studies. First, so-called "pure" Silence is relatively easy for automatic speech recognition systems to detect. Since daylong recordings with infants and toddlers will contain long stretches of Silence during naptimes, one should be mindful how these sections are incorporated, as they could inflate accuracy estimates. Second, LENA segmentation is meant to identify when humans are talking — from the beginnings to the ends of their utterances. While spontaneous speech typically includes pauses or Silence — and LENA modeling certainly recognizes such Silence — LENA algorithms also incorporate a minimum duration criterion of 800ms for labeling Silence segments. Thus, if a coder hears a bit of Silence within an adult or child segment that is shorter than 800ms, it should NOT be counted as an error. Similarly, human segmentation label categories have their own minimum duration requirements and sometimes may include Silence to achieve them. (See Appendix A for more information on minimum duration criteria.)

PRIMARY & SECONDARY SEGMENTATION LABELING

Accuracy for Key Child segmentation labeling (a primary label) was a high priority from the outset of algorithm development. A crucial component of that effort was to avoid confusion between child and adult segments so the child would not appear more linguistically sophisticated than he/she was. Further, since adult male and female segments contribute to the AWC and CTC metrics, LENA also prioritized distinguishing adult voices from other sounds, especially TV/Electronic Sounds. However, since the four secondary segmentation labels — Other Child, Overlap, Noise, and Silence — were unutilized in the core reports, LENA's developers were less concerned with accurately distinguishing between them. In general, it is therefore not advisable to merge primary with secondary segmentation labels in validation analyses. Although it may seem intuitive to combine Key Child and Other Child, for example, they were modeled from the outset with very different goals in mind, and only Key Child segments are processed to identify Child Vocalizations.

FRAME-LEVEL CODING

Although most attempts to validate LENA accuracy are conducted at a macro level, it may be helpful to consider briefly how LENA segmentation and labeling is achieved. Audio recordings are first examined at a *frame* level encompassing only 10ms of audio. A preliminary label for the frame is set by comparing its acoustic properties statistically to the eight pre-defined segmentation category models. A recording-level best-fit solution is then generated, utilizing minimum duration constraints (ranging from 600 to 1000ms) to combine frames into fixed boundary segments. An adult segment, e.g., may then comprise not only speech sounds but also Silence. LENA users cannot access frame-level coding, so it is recommended that the LENA segment be the smallest unit of analysis for validation efforts, even given more granular human coding. Finally, because segmentation boundaries will almost certainly differ between humans and LENA, we recommend counting reasonably overlapping label agreement as a "hit."

Regarding word and vocalization counts, say 80% of an adult LENA segment overlaps with a human-defined adult segment and 20% with an adjoining Silence segment. It seems reasonable then to assign 80% of the LENA word count to the matching segment and 20% to the mismatched one. However, doing so implicitly assumes that the adult speech is spread uniformly across the segment, which of course is not necessarily or even usually the case. It may well be that 100% of the detected speech is contained in the overlapping section (i.e., where machine-coded and human-coded labels agree), in which case the estimated accuracy of the count may have been unfairly reduced. And as previously mentioned, the word

count estimate itself incorporates the full segment duration into its calculation. Thus, careful thought should be given to choices made when comparing LENA-generated to human-generated counts. In general, we recommend evaluating counts at a more macro level. For example, if 5-minute sections of recording were transcribed and counts generated, then the unit of analysis would be the sum total LENA and transcription counts for the 5-minute section, not the LENA segment or human-coded utterance.

PRECISION VS. SENSITIVITY

A further consideration for LENA's developers was weighing the loss of "good" or potentially usable data against the consequence of mislabeling. Discarding data may seem very counterintuitive to those used to more traditional, labor-intensive methods of data collection. But since LENA's automated approach has the advantage of recording continuously throughout the day and thus samples from a relative abundance of data, generating accurate primary segmentation labels was deemed more important than losing some small percentage of data. It follows that those doing validation work should not expect sensitivity (e.g., amount of human-identified Key Child matching LENA coding) to be higher than precision (e.g., amount of LENA-identified Key Child matching human coding). In other words, by design, LENA can be expected to tend toward undercounts when compared to human transcription. However, across recordings and depending on the setting and other factors, it can be expected that LENA counts should correlate reasonably well with transcriber counts, even when differences in absolute counts exist.

INTER-RATER RELIABILITY

It is standard practice to measure inter-rater reliability by having all human raters code the same audio segments and then compare agreement among them. Given the inherent subjectivity of human judgments, there is never perfect agreement between coders. Respectable agreement using a Cohen's kappa is only around 80%. It is therefore very important that inter-rater reliability always be included when reporting on LENA accuracy, bearing in mind that error rates between human coders and LENA cannot be expected to be better than those between human coders, upon which LENA modeling itself was based.

ASSESSING CONVERSATIONAL TURNS ACCURACY

The LENA Conversational Turns metric is operationally defined as alternations between Key Child[3] and Adult Male/Female segments that include speech-related vocalizing and are separated by no more than five seconds of Silence or other nonspeech. A given segment can count toward only one turn. (See Appendix C for more information about how Conversational Turns are counted.) In one sense then, if the goal of validation is to report accuracy for Conversational Turns given what they were designed to represent, no additional coding is needed beyond human segment labeling. (That is, one can simply apply LENA's rule set to the human-labeled data.) However, if the goal is to compare LENA turn counts to frequency of child-caregiver exchanges involving child- and adult-directed speech, then such content information must be coded. It should be noted here that LENA's automated labeling approach cannot and does not claim to identify the directionality of any speech, so validating turns based on directedness exceeds the parameters of the technology's capabilities. See Appendix C for more detail on turns.

OUTLIER ANALYSES

When analyzing segmentation accuracy from a relatively small quantity of coded audio, it is important to carefully consider the influence of outliers, which can disproportionately influence results. Evaluators are encouraged to reference Aduinis, Gottfedson & Joo (2013) for best practices in working with datasets including extreme outliers. When identified, results should be reported both with and without outliers, regardless of how different the outcomes might be. We also recommend that researchers listen to outlier audio sections when available and report circumstances that may lead to increased error rates, as this will allow LENA users to gauge the probability of such errors occurring in their own recording data.

[3] More specifically, Key Child segments eligible to be part of a Conversational Turn must include speech-related communicative vocalizing; segments containing only vegetative sounds (e.g., breath, burp) or fixed signals (e.g., cries, screams) do not contribute to LENA turns.

APPENDIX A: OPERATIONAL DEFINITIONS FOR AUTOMATED SEGMENTATION LABELS AND LENA METRICS

This document details each of LENA's automated segmentation labels and core report metrics and is intended to clarify the technology's design and purpose. A clear understanding of these definitions is crucial when assessing LENA performance and when designing validation studies.

LENA's eight segmentation labels (described below) can be sorted into four primary and four secondary categories. The four primary segmentation labels (Key Child, Adult Female, Adult Male, TV/Electronic Sounds) contribute directly to the core LENA report metrics: Child Vocalizations, Adult Word Count, Conversational Turns Count, and TV/Electronic Sounds duration. The four secondary segmentation labels (Other Child, Overlap, Noise, and Silence) are identified but unutilized — that is, they do not contribute to LENA reports and are, for practical purposes, eliminated from further analyses.

**Note: Recording segments assigned to primary and secondary segmentation labels other than Silence are then compared statistically to the acoustic model for Silence. When this comparison exceeds a certain threshold, the segment is designated 'faint' (f). Faint segments other than Key Child do not contribute to report metrics.

Primary Segmentation Labels

Key Child: Includes any sounds originating from the mouth of the child wearing the LENA recorder, including speech-related babbles, words, and sentences, as well as vegetative sounds (e.g., burps, breaths) and fixed signals (e.g., cries, screams, laughs). Key Child segments are automated representations of child utterances and have a minimum duration of 600ms. Transcription data from children between 2 months and 48 months of age were used to create the acoustic models for labeling Key Child. These models are optimized to identify children in this age range who are wearing the LENA recorder (i.e., whose mouths are within a few inches of the microphone).

Adult Female: Includes post-pubescent female voices. Adult female segments are automated analogs of Female Adult utterances with a minimum duration of 1000ms (1 second).

Adult Male: Includes post-pubescent male voices. Adult male segments are automated analogs of Male Adult utterances with a minimum duration of 1000ms (1 second).

TV/Electronic Sounds: Includes any sound emanating from an electronic speaker, e.g., from radio, television, or electronic toys. These segments have a minimum duration of 1000ms (1 second).

Secondary Segmentation Labels

Other Child: Includes vocalizing from male and female pre-pubescent children in the immediate vicinity (within 6 to 10 feet) of the Key Child. These segments have a minimum duration of 800ms. Note that vocalizations of older/post-pubescent children are less likely to receive this label.

Noise: Includes ambient environment sounds, from short bumps to long rattles to persistent white or pink noise (e.g., a loud generator or fan close by), that are unrelated to human vocalization and do not originate from TV or other electronic sources. These segments have a minimum duration of 800ms.

Overlap: Includes human vocalizing detected contemporaneously with other environmental human or nonhuman sounds (e.g., human+human or human+noise) with a minimum duration of 800ms.

Silence: Includes recording periods of at least 800ms minimum duration with little or no acoustic content or for which the acoustic energy level is at or below 32 decibels. Note that in a natural recording environment, periods of "true" Silence may be rare, and the LENA recorder's very sensitive microphone will register even very faint or distant sounds.

LENA Core Report Metrics

1. **Adult Word Count:** Estimate of the number of words spoken by post-pubescent males and females in the child's environment. LENA algorithms do not identify words or recognize their semantic content; instead, they generate an unbiased word count estimate for each adult female or male segment using acoustic information in the segment, specifically vowel and consonant distribution and durations, as well as length of utterance.

2. **Child Vocalization Count:** Estimate of the number of times the child produced communicative (i.e., speech-related) vocalizations, NOT including vegetative sounds (sounds related to respiration or digestion) or fixed signals (instinctive reactions to the environment such as cries). Child Vocalization counts are generated only for Key Child segments. A Child Vocalization has no maximum duration, but one is distinct from an additional vocalization when the two are separated by at least 300 milliseconds of Silence or other sounds. For example, the babble "bababa" and the sentence "I want my num num" would each count as one vocalization as long as any within-vocalization pauses do not exceed 300ms.

3. **Conversational Turns Count:** Estimate of the number of alternations between the Key Child and an adult in his/her environment. The Conversational Turns metric is calculated exclusively between segments labeled Key Child (including a vocalization) and segments identified as Adult Female or Adult Male (including a word) that are separated by no more than 5 seconds of Silence or other sounds. For example, if the Key Child produces a speech-related vocalization and an adult responds within five seconds, that would be counted as one turn. Similarly, a turn is counted when an adult says something, and the child responds within five seconds. Please see Appendix C for more specific details on LENA turn counting in automatically segmented files.

4. **TV/Electronic Sounds:** Estimate of the amount of time that sounds originating from an electronic speaker were dominant in the child's environment. That is, this metric reflects the total duration of segments labeled TV/Electronic Sounds.

Note that LENA's segmentation and labeling process is designed to identify the dominant sound source in a child's environment discretely over very short periods of time, not to identify events or conditions that may occur over a longer span of time. For example, say a child and adult are interacting for seven minutes while a nearby TV is on: a transcriber might note the TV duration as the entire seven minutes, but LENA will only sum those periods when the TV sound is identified as dominant. Thus, the LENA segmentation of that interaction could include a combination of Adult, Key Child, TV, Overlap, and Silence segments which, when added together, equal seven minutes total duration.

APPENDIX B: LENA AUTOMATED SEGMENTATION PERFORMANCE

This document reports sensitivity and precision statistics for LENA's eight primary and secondary segmentation labels. Performance results are based on 82 hours of human coding from 94 daylong audio recordings completed by families with children 2 months – 48 months of age. (For a description of the coding process see LENA Technical Report LTR-06-2.[1])

LENA segmentation labels are described in detail in Appendix A. Briefly, the four primary LENA labels contributing to report metrics are: Key Child (CHN), Adult Female (FAN), Adult Male (MAN), and TV/Electronic Sounds (TVN). Primary segmentation labels and percent agreement are highlighted in purple in Tables 1-3. In addition to primary and secondary segmentation labels, these tables include an Other category (OTH), into which all other types of sounds were grouped. This category for the rows (human transcribers) and columns (LENA automated segmentation) is defined as follows:

Other (OTH) by-row: Includes anything human transcribers labeled unclear, including faint sounds and segments they could not attribute to a specific speaker or other sound source with confidence. This category also includes cases of human-identified Overlap with adult vegetative sounds and Overlap when the Key Child was not wearing the vest (instances of which were both very rare).

Other (OTH) by-column: Includes anything LENA automatically identified as unclear (i.e., includes all faint/far labels).

Red highlighting in the tables denotes miscategorization that could erroneously *inflate* LENA estimates for Adult Word Count, Conversational Turns Count, Child Vocalization Count, and TV/Electronic Sounds reports. Yellow highlighting represents errors whereby something that should have contributed to reports was eliminated, which could result in *decreases* in these LENA report estimates.

[1] https://www.lena.org/wp-content/uploads/2016/07/LTR-06-2_Transcription.pdf

Table 1 details sensitivity for each of LENA's eight segmentation labels (plus Other), derived from frame-level analysis and shown as percentages. Each row sums to 100 percent of what the transcribers labeled in each category. For example, the first row shows that LENA correctly labeled 67 percent of what the human coders identified as Key Child, while 7% of real Key Child speech was erroneously labeled as adult female (inflating word counts) and 13% was labeled as Other Child and eliminated from analysis.

Table 1. LENA Segmentation Sensitivity

		LENA SEGMENTATION									
		CHN	FAN	MAN	TVN	CXN	NON	SIL	OTH	OLN	TOTAL
HUMAN CODING	CHN	**67%**	7%	0%	0%	13%	0%	4%	6%	4%	100%
	FAN	2%	**74%**	7%	3%	5%	0%	2%	6%	2%	100%
	MAN	1%	10%	**72%**	6%	1%	0%	2%	7%	2%	100%
	TVN	2%	6%	5%	**61%**	4%	2%	6%	11%	4%	100%
	CXN	7%	14%	0%	1%	**64%**	0%	2%	6%	5%	100%
	NON	2%	3%	3%	7%	1%	**3%**	13%	56%	12%	100%
	SIL	1%	3%	1%	2%	1%	0%	**80%**	11%	0%	100%
	OTH	4%	6%	3%	6%	9%	0%	19%	**46%**	6%	100%
	OLN	8%	19%	11%	8%	8%	0%	1%	15%	**30%**	100%

LENA's automated processing can identify *when* speakers are overlapping but not *which* speakers are included in the overlap. Thus, OLN "error" is considered less problematic since it could well include the speaker the transcriber identified. For example, 19% of what the human listener called Overlap was labeled as Female Adult by LENA. Although the human listener identified two competing sounds, the LENA segmentation algorithm recognized Female Adult as dominant. Of course, it is also possible that the Overlap did not include a Female Adult, thus reflecting a true error.

Yellow shading indicates situations in which true adult and child speech was assigned to a secondary segmentation label and eliminated from analysis, representing a loss of data and potentially leading to undercounts in the LENA reports. Table 2 below collapses the secondary segmentation label errors (yellow) so it is easier to see the percent of useful data that was erroneously eliminated.

Table 2. LENA Segmentation Sensitivity, Collapsing Secondary Speaker Labels

		LENA SEGMENTATION				
		Child	Adult	TV/Media	Secondary	Total
HUMAN CODING	Child	66.5%	7.4%	0.2%	25.9%	100.0%
	Adult	2.0%	**80.7%**	3.5%	13.8%	100.0%
	TV/Media	1.8%	10.8%	**60.6%**	26.8%	100.0%
	Other	3.7%	12.4%	5.5%	**78.3%**	100.0%

As Table 2 shows, 25.9% of the Key Child's vocal output was assigned a secondary label and eliminated. As mentioned in the main text, the algorithms were designed to prioritize elimination of good data to minimize the amount of miscategorization in the LENA report metrics.

Table 3 details LENA's precision performance, showing the amount and type of error in LENA-identified segments. The breakout of automated LENA labels into the corresponding transcriber labels is shown as columnar data summing to 100 percent. Looking at the first column, we see that 75% of what LENA labeled as Key Child the human coder also labeled Key Child, while 4% was actually Female Adult and 1% was Male Adult, inflating Child Vocalization estimates.

Table 3. LENA Segmentation Precision

		LENA SEGMENTATION								
		CHN	FAN	MAN	TVN	CXN	NON	SIL	OTH	OLN
HUMAN CODING	CHN	**75%**	4%	0%	0%	24%	4%	3%	3%	6%
	FAN	4%	**67%**	14%	8%	13%	1%	2%	5%	5%
	MAN	1%	3%	**52%**	5%	1%	1%	1%	2%	1%
	TVN	1%	1%	2%	**34%**	2%	8%	1%	2%	1%
	CXN	2%	2%	0%	1%	**27%**	1%	0%	1%	2%
	NON	3%	3%	7%	22%	3%	**71%**	16%	54%	28%
	SIL	2%	2%	2%	3%	3%	5%	**68%**	8%	0%
	OTH	2%	2%	3%	6%	10%	4%	8%	**15%**	5%
	OLN	11%	15%	20%	20%	18%	6%	1%	11%	**51%**
	TOTAL	100%	100%	100%	100%	100%	100%	100%	100%	100%

APPENDIX C: LENA RULES FOR COUNTING CONVERSATIONAL TURNS

LENA Conversational Turns can be defined simply as alternations between adult and Key Child voices. However, there are many situations that can occur in a natural, spontaneous speech environment in which LENA technology must systematically apply certain rules. The purpose of this document is to elucidate the process implemented to identify Conversational Turns in LENA's content-free approach, which does not involve recognizing words, assessing semantic content, or identifying directionality. This information should be considered carefully when assessing Conversational Turns performance.

Table 1 lists the LENA segmentation labels available in LENA's automated processing exports.[1] Note that here segments closely matching LENA models (i.e., "near and clear") are provided individually, whereas far-field, faint, and/or unclear sound source labels are collectively shown as FUZ. For a complete listing of labels in LENA files, see Table 1 in LENA Technical Report LTR-04: *The LENA™ Language Environment Analysis System: The Interpreted Time Segments (ITS) File.*[2]

Table 1. LENA Segmentation Labels

Source of Sound	LENA Label
Key Child	CHN
Female Adult	FAN
Male Adult	MAN
Other Child	CXN
Overlapping Sounds	OLN
Noise	NON
TV/Electronic Sounds	TVN
Silence	SIL
Uncertain/faint	FUZ

Necessary Sound Segments

There are two initial considerations to qualify segments for a turn. First, turns must include both a Key Child and an adult male or female segment. Second, Key Child segments must include at least one speech-related vocalization,[3] and adult segments must include at least one word. If an adult or Key Child segment includes only cries or vegetative sounds, it cannot contribute to a turn.

[1] Table 1 labels are the same as those referenced by ADEX.

[2] https://www.lena.org/wp-content/uploads/2016/07/LTR-04-2_ITS_File.pdf

[3] See Appendix A for a detailed definition of Child Vocalizations.

Barriers to Turns

Alternations between Key Child and an adult that otherwise would be called turns can be "interrupted" by the presence of another child. When this happens, the turn is precluded. For example, if the Key Child vocalizes and then another child in the vicinity speaks and is followed by a Female Adult, no turn is counted because it is not clear whether the adult was responding to the Key Child or to the Other Child.

Barrier to turns: Other Child – CXN

Example 1: CHN – CXN – FAN → 0 turns

Similarly, speech from an adult speaker of a different gender can interrupt a turn, in which case the second adult simply becomes the new turn initiator/responder. (Speech from a same gender adult that spans multiple segments is allowed, but a turn is only counted for the segment nearest the child's.)

Barrier to turns: Different gender adult segment

Example 2: FAN – MAN – CHN → 1 turn between MAN and CHN

Note that LENA export segmentation mapping (*.its files) only counts a turn on the response segment, though the initiating segment is also tagged.

Allowable Intervening Segments

Any other type of segment or group of segments may intervene between a qualifying Key Child and adult segment, provided the time between the initiation and response remains under 5 seconds. Allowable intervening segments include Overlap, Noise, TV/Electronic Sounds, Silence, and unclear/faint sounds (FUZ).

Segments that may come between adult and child within a turn: OLN, NON, TVN, SIL, FUZ

Example 3: CHN – OLN – NON – SIL – MAN → 1 turn, if the total intervening duration is < 5 seconds

Note: Key child segments with zero vocalizations and adult segments with zero words are not barriers.

Counting Conversational Turns

A Conversational Turn must include one initiation and response. A response already included in one turn may not serve as the initiation of another turn. In other words, once a segment has been "used" it cannot contribute to a new turn.

Example 4: FAN – CHN – FAN → 1 turn

Example 5: FAN – CHN – FAN – CHN → 2 turns

Directionality

Since LENA does not identify speech content, Conversational Turns may derive from adult utterances that are not actually directed to the Key Child. For example, when a parent is talking on the phone and holding a baby engaged in vocal play, adult-child alternations that meet the usual conditions will be counted as Turns. To evaluate the extent to which Conversational Turns include child-directed speech will require an additional layer of human coding to identify utterance content.

Summary

In sum, a LENA Conversational Turn must contain at minimum one Key Child segment with a vocalization and one adult segment with a word. A LENA turn has one initiation and one (and only one) response. The response must occur within five seconds of the initiation, and the initiation-response interim can include any sound segments except from another child or a different gender adult. To match LENA's approach to turn counting, one can simply look for adult or child initiations and a response within 5 seconds, without another child segment intervening. The directionality of adult speech within turns may additionally be coded, but any evaluation of the accuracy of Conversational Turns should note that LENA cannot, and has never purported to, differentiate adult-directed from child-directed speech.