

A web-derived typology of multimodal infant crying: Taxonomy, a text-only web-corpus comparison, and ethical issues

Jacob Rigal

Université Paris Cité / LLF, CNRS
jacob.rigal@u-paris.fr

Jonathan Ginzburg

Université Paris Cité / LLF, CNRS
yonatan.ginzburg@u-paris.fr

Abstract

Web video platforms contain naturally occurring audiovisual interaction, but their use as corpora requires explicit decisions about selection, metadata, annotation, privacy, and redistribution. We present a web-derived data set of infant cry contexts coding 250 publicly accessible videos from YouTube, Instagram, and TikTok, covering vaccination, discomfort, gas, fussiness, colic-like crying, hunger, birth, sleepiness, overstimulation, and other pain contexts. Candidate videos were labeled from audiovisual evidence; validation currently covers 230 videos (92.0%), of which 226 receive a current taxonomy label and enter the audiovisual distributional tables. Validation also served to refine the taxonomy, revealing systematic under-recognition of sleep-related visual cues and recurrent subtypes within the broader *Fussiness* category. To test what can be recovered from linguistic material alone, we contextually classified 4,600 *baby* + *CRY* concordance lines from enTenTen21 with OpenAI GPT-5.6 Sol, recording a category, confidence level, and explanation for each line. A human audit of the first 460 lines (10.0%) covered all output categories and yielded 91.5% normalized categorical agreement ($\kappa = .885$; bootstrap 95% CI [.849, .917]). After audited corrections, 72.2% of all concordance lines were *Non-target* or *Unclassifiable*. This shows the limits posed by a lack of multi-modal context in text-only specialized corpora. Finally, we illustrate a proposed enriched transcription format for representing infant vocal activity, visible behavior, adult speech, caregiving action, and procedural context. We propose a privacy-preserving, source-decoupled derivative resource whose release would be subject to institutional review and platform/licensing constraints.

1 Introduction

The web has been used as a source for linguistic corpora for a number of years (Kilgariff and

Grefenstette, 2003; Baroni et al., 2006). More recently, web video has become a major source for audiovisual datasets: AudioSet uses human-labeled YouTube clips for sound-event recognition (Gemmeke et al., 2017). Elsewhere, YouTube-8M provides large-scale video labels and features (Abu-El-Haija et al., 2016), while VGGSound links audio classes to visible sound sources in YouTube clips (Chen et al., 2020). Another set, VoxCeleb, draws speaker data from online video (Nagrani et al., 2017), and still another, MOSI and MOSEI, use online opinion videos for multimodal sentiment and emotion analysis (Zadeh et al., 2016, 2018). Action-recognition datasets such as Kinetics and AVA show how web or already-circulating video can support fine-grained temporal annotation (Kay et al., 2017; Gu et al., 2018). These resources demonstrate the value of web audiovisual data, but they also foreground problems central to WaC research: platform bias, metadata sparsity, label noise, unstable availability, and the difficulty of documenting how source media become datasets. Infant-crying videos make these problems especially concrete because they combine highly variable user-generated recordings with sparse metadata, sensitive family and medical contexts, and identifiable minors.

A central difficulty in infant-cry research is that the cause of crying is often underspecified by the cry signal itself. Although cries reliably convey distress, specific eliciting causes are considerably harder to recover from acoustics alone. For example, Lockhart-Bouron et al. found that neither trained adult listeners nor machine-learning models could reliably distinguish hunger, discomfort, and isolation from cry acoustics; in their longitudinal corpus, cry causes were instead labeled primarily from the caregiver action that successfully stopped the crying, which agreed with the caregiver’s prior attribution in 75% of cases (Lockhart-Bouron et al., 2023). This illustrates an important property of

infant crying as interactional data: the eliciting context may be ambiguous at cry onset and may become clearer only through the surrounding situation and the caregiver's subsequent response. Feeding, repositioning, restoring proximity, removing a source of discomfort, or terminating an aversive stimulus can provide evidence that was not available from the cry alone.

At the same time, there is no broadly established data-driven typology spanning the range of everyday contexts in which young infants cry. Existing corpora use heterogeneous and often relatively narrow label sets, ranging from hunger and discomfort to clinically defined categories such as asphyxia or neurological conditions (Ji et al., 2021; Laguna et al., 2023). Reviews of infant-cry research also emphasize the scarcity of well-documented labeled data: many datasets have been collected in hospitals, NICUs, clinics, or controlled home studies, and many remain private or available only through restricted research access (Ji et al., 2021). This is particularly consequential in early infancy, where naturalistic, context-rich recordings are difficult to collect at scale. Our aim is therefore to develop an empirically grounded, context-sensitive, multimodal taxonomy of infant crying episodes from naturally occurring audiovisual material rather than to assume that discrete cry types can be read directly from acoustics alone. Web video is useful for this purpose because it can preserve the sequence linking an antecedent event or state, the infant's vocal and visible behavior, caregiver interpretation and action, and the subsequent continuation or resolution of distress. The resulting taxonomy is thus a typology of recoverable eliciting contexts, cues, and interactional evidence.

The contribution of this paper, in addition to an operational taxonomy of infant crying, is a compact data-centric description of how online audiovisual data can be used for empirical web-as-corpus research. We describe the analyzed sample, coding scheme, quality constraints, and validation procedure, together with a proposed privacy-preserving reuse model submitted for institutional ethical and data-protection review.

2 Corpus construction

Links to candidate videos were located between March and July 2026 on YouTube, Instagram, and TikTok. Undergraduate students contributed only to locating public links to videos of infants or ba-

bies crying and excluding links clearly concerning older children; they did not code or analyze the videos. No fixed search-term list or standardized retrieval procedure was imposed, and link-finding strategies could vary. The aim was broad, non-targeted discovery rather than targeted sampling of predefined cry contexts. Because retrieval was neither standardized nor randomized, the resulting sample should not be interpreted as a probability sample of infant-crying videos on these platforms.

The initial link set contained 325 candidate video records. Of these, 47 are marked as duplicates in the corpus table, and 28 additional non-duplicate records do not receive a retained cry-context label. Removing these 75 records leaves 250 unique classified videos. Video coding was carried out separately from link discovery. Initial context labels were assigned from the audiovisual record by the research team; Hexiang Wu, a graduate student in linguistics, also coded videos. The first author subsequently conducted an audit/validation pass in which the audiovisual context was re-examined and the original label could be confirmed, revised, or qualified with an explanatory note. This audit was not designed as a blinded, fixed-codebook inter-annotator agreement experiment: disagreements were also used to refine the coding guide. The 45 vaccination cases underwent a separate video-level context check and were all confirmed. Uploader-provided titles and descriptions were used for discovery but not as ground-truth labels. Corpus-level metadata used in the present analysis are deliberately sparse, containing context label and video duration; validation judgments are retained as a separate audit layer.

Compilations were retained when their crying episodes consistently represented the same cry context throughout; in such cases, the video received that single context label. Compilations for which a consistent cry context could not be established were not assigned a single-category label. The label *Multiple* was reserved for retained items in which more than one distinct cry context was independently supported. Table 1 summarizes the selection criteria.

The analyzed sample follows documentation principles developed for NLP and ML (machine learning) datasets (Bender and Friedman, 2018; Gebu et al., 2021; Paullada et al., 2021). For this domain, documentation must also encode what is deliberately *not* released and what metadata were *not* used. This is particularly important for web-

Criterion	Operational decision
Search protocol	Undergraduate students located public video links using self-selected search strategies; they did not code the videos. Older children were excluded.
Infant cry present	At least one audible infant cry or cry-like distress episode.
Context recoverable	Eliciting context visible/audible or explicitly attributed within the recording.
Audio usable	Cry pulses and adult speech audible enough for coding.
Video usable	Relevant visual context sufficiently observable for context coding.
Metadata used	Corpus-level analysis uses validated context label and video duration.
Sensitive content minimized	Source videos are not part of the proposed reusable corpus.

Table 1: Selection criteria for the web-derived infant crying corpus.

Field group	Examples
Corpus-level meta-data	validated context label; video duration.
Discovery information	public links located through web search, not used as analytic labels.
Uploader meta-data	titles/descriptions not used as ground-truth labels or analysis variables.
Proposed enriched transcript	infant cry notation; visible infant behavior; caregiving actions; speaker-attributed adult speech; salient scene/procedural events.
Proposed reuse	subject to ethical approval; de-identified, source-decoupled transcripts and annotations.

Table 2: Metadata and annotation fields used or proposed in the present corpus design.

derived audiovisual data, where large-scale collection practices have been criticized for weak consent, poor filtering, and opaque downstream reuse (Birhane and Prabhu, 2021). We therefore distinguish three layers: web-based discovery, audiovisual evidence used for annotation, and source-decoupled linguistic or multimodal representations proposed for reuse. In the current version, corpus-level metadata are restricted to context label and duration.

A further design decision is to keep the unit of discovery distinct from the unit of analysis. The web item is a video, but the linguistic or interactional evidence may be an infant vocalization, a caregiving action, an adult utterance, a visible infant behavior, or a procedural event such as an injection. This distinction makes it possible to use video-level labels for broad distributional work while representing selected interactions more richly in a source-decoupled transcript. The same design also avoids over-interpreting titles, uploader descriptions, or platform categories as linguistic

metadata; these were discovery aids, not coding evidence. Platform of origin was not retained as an analytic variable in the sparse corpus-level table, so the present study does not report platform-stratified counts or treat platform as an explanatory factor.

3 Composition and taxonomy

The analyzed sample contains 250 unique classified videos. Current validation covers 230 (92.0%). Of these, 226 receive one of the current taxonomy labels and are summarized in Tables 3 and 4; four validation-covered records do not receive a current taxonomy label (three contain no crying and one remains unresolved between *Discomfort* and *Overstimulation*). Two validation judgments coded as proximity-seeking are treated as a descriptive subtype of *Fussiness*, consistent with the taxonomy below. Within the 226 labeled cases, 20 are *Unknown* and three are *Multiple*, leaving 203 substantive single-category audiovisual cases for the cross-modal comparison in Table 6. Labels describe the *eliciting context* supported by the recording, not an acoustic diagnosis. The category *Colic-like* is therefore used for audiovisual data, while *Colic* is retained below for web-text passages that explicitly use that term. The initial cry-context taxonomy was developed from existing research on infant crying, including work on acute pain, hunger, fussiness, discomfort, colic, and related behavioral states (Bellieni et al., 2004; Goberman and Whitfield, 2013; Parga et al., 2020; Laguna et al., 2023; Hammoud et al., 2024; Ji et al., 2021). It was subsequently refined through validation of the audiovisual labels: recurrent ambiguities and boundary cases motivated more explicit assignment criteria, including the use of *Colic-like* rather than *Colic* for audiovisual cases and stricter distinctions among *Gas*, *Discomfort*, *Fussiness*, *Birth*, and *Unknown*. Real-world cry detection likewise shows substantial in-the-wild noise and generalization challenges (Yao et al., 2020).

Existing infant-cry resources have largely been designed for acoustic classification. The Baby Chillanto database, for example, contains short cry segments grouped into clinically or behaviorally defined categories, while the crowdsourced Donata-Cry corpus contains parent-contributed audio recordings with contributor-supplied reason labels (Ji et al., 2021). The present resource differs in retaining the naturalistic audiovisual context surrounding crying episodes and in labeling the elicited

Cry context	<i>n</i>	Mean length
Vaccination	45	1:43
Discomfort	44	2:16
Sleepiness	34	1:00
Fussiness	30	1:27
Hunger	13	1:27
Colic-like	12	2:05
Birth	8	3:46
Other Pain	8	2:08
Overstimulation	6	1:50
Gas	3	1:31
Multiple	3	10:45
Unknown	20	1:06
Total	226	1:50

Table 3: Validation-covered audiovisual cases assigned to the current cry-context taxonomy ($n = 226$).

Functional group	Corpus labels included	<i>n</i>
Homeostatic need	Hunger; Sleepiness	47
Immediate acute nociceptive	Vaccination; Other Pain	53
Externally elicited discomfort	Discomfort	44
Internal gastrointestinal / regulatory	Colic-like; Gas	15
Regulatory / interactional distress	Fussiness; Overstimulation	36
Perinatal transition	Birth	8
Unassigned / mixed context	Multiple; Unknown	23
Total		226

Table 4: Functional grouping of validated audiovisual cry-context labels.

ing context from multimodal evidence rather than treating isolated cry acoustics as the primary unit of analysis.

3.1 Validation-driven taxonomy refinement

Validation functioned not only as a reliability check but also as a mechanism for revising the coding scheme. The clearest case concerned *Sleepiness*. Among 40 videos initially labeled *Unknown* that received a validation judgment, 28 (70%) were re-assigned to *Sleepiness*. This pattern exposed an omission in the first version of the taxonomy: visually recoverable sleep-state cues had not been operationalized sufficiently. The refined guide therefore treats a sleep transition, settling sequence, or partial-arousal state as relevant when supported by converging behavioral evidence such as repeated or prolonged eye closure, slow eyelid opening and closing, yawning, eye rubbing, or drowsy re-arousal. Scene context such as a crib or bed can support the interpretation but is not sufficient by itself.

Validation also showed that *Fussiness* covers

several recurrent forms of regulatory or protest distress rather than a single eliciting situation. We therefore use the label operationally, not as a claim that the crying has no real cause, is trivial, is under voluntary control, or ought to be ignored. Cry-fuss behavior can arise from interacting biological, environmental, and psychosocial factors, and the same overt crying may organize different caregiver responses (Douglas and Hill, 2013). Within the present material, recurrent subtypes include *goal-blockage protest*, for example when an object is removed or an ongoing activity ends; *proximity-seeking or separation-related protest*, when access to a caregiver is interrupted; and *socially elicited distress*, including contagious crying following another infant’s cry. These remain descriptive subtypes rather than new top-level labels because the available evidence does not yet support stable independent categories.

This caution is especially important for sleep-related crying and so-called *cry-it-out* practices. Behavioral sleep interventions are heterogeneous, and evidence varies with age and method: trials of graduated-extinction approaches in older infants have not detected adverse effects on attachment or later emotional/behavioral outcomes, whereas work focused on the first six months has questioned the benefits of broadly applying such interventions and has emphasized the complexity of cry-fuss regulation (Gradisar et al., 2016; Price et al., 2012; Douglas and Hill, 2013). We therefore do not treat sleep training itself as evidence of neglect or trauma, nor do we equate any bedtime cry with *Sleepiness* or *Fussiness* in the audiovisual guide. Caregiver non-response, delayed response, or withdrawal may instead be represented as a separate interactional property when it is observable.

Qualitative inspection also surfaced a small set of contexts not well captured by the present top-level labels, including abandonment, prolonged unattended crying, frightening or rough caregiving, forced displacement, and distress following potentially traumatic events. Such cases motivate future distinctions concerning *caregiver responsiveness/adverse caregiving context*, but we avoid assigning clinical or legal labels such as *trauma* or *neglect* from a short web video alone. This distinction matters because adverse caregiving encompasses different forms of threat and deprivation that should not be collapsed into a single construct (Guyon-Harris et al., 2021); infant crying can also be a trigger for harmful caregiver responses,

making the caregiver response analytically distinct from the original cause of crying (Reijneveld et al., 2004). A further exploratory dimension concerns event continuity: many *Fussiness* cases involve ordinary but unwanted limitations on an ongoing goal or social-regulatory state, whereas *Overstimulation* more often follows abrupt, intense, novel, or expectancy-disrupting input. We treat these distinctions as hypotheses for future coding rather than additional current labels. Because the audit and taxonomy refinement were deliberately intertwined, the disagreement patterns remain evidence for codebook development rather than a fixed-codebook reliability test.

4 Web-text comparison: retrieval and cry-context recoverability

The audiovisual labels above are assigned from what can be seen and heard in each video. To test how far the same taxonomy can be recovered from linguistic material alone, we used the 52,268,286,493-word English enTenTen21 web corpus through Sketch Engine (Kilgariff et al., 2014). We queried the corpus with the CQL expression `[lemma="baby"] [lemma="cry"]`, which returned 14,046 concordance hits. No additional filters, sorting parameters, or randomization were applied. Sketch Engine therefore displayed the concordances in its default corpus order, and the present analysis uses the first 4,600 exported lines. The sample is consequently a contiguous, non-random subset of the query results. It is not a prevalence estimate, but a test of how often local web text supplies enough context to classify an infant-crying mention.

4.1 Contextual classification and validation

Each concordance was read from the exported local KWIC context (*Left* + *KWIC* + *Right*); no extended document context was consulted. Each line was judged individually with OpenAI GPT-5.6 Sol in ChatGPT. The procedure was deliberately contextual rather than keyword-rule based: the model determined whether an actual infant-crying event was described and, when possible, which eliciting context was supported by the passage. The output fields were Classification, Confidence (High/Medium/Low), and Contextual_basis, a short explanation of the evidence for the decision. The label set contained the ten shared cry contexts plus *Multiple*, *Unclassifiable*, and *Non-target*. *Mul-*

tipple was used only when more than one context was independently supported; *Unclassifiable* when an infant cry was present but its cause was not recoverable; and *Non-target* for figurative, quoted, imagined, fictional, device-related, lyric, or other non-event uses. Classification was performed interactively under a stable decision protocol rather than as a single one-shot API prompt; Appendix B records the operative rules.

Across the 4,600 lines, confidence was High for 4,094, Medium for 505, and Low for one. A human audit covered lines 1–460, exactly 10.0% of the classified set. The block was contiguous rather than randomly sampled, and the audit was corrective rather than blind: the auditor reviewed each KWIC line together with the assigned label, leaving Validation blank when the assignment was accepted and entering a correction or note otherwise. Of the 460 lines, 419 were accepted without an entry and 41 received a non-empty validation entry. For comparison with the declared label inventory, schema-equivalent wording was normalized (for example, *Unknown* to *Unclassifiable*); purely descriptive validator notes that did not constitute a taxonomy label were likewise collapsed to *Unclassifiable*. Two such entries were therefore equivalent to the original category, giving 421/460 normalized categorical agreements (91.5%; Wilson 95% CI [88.6, 93.7]). Cohen’s κ was .885 (10,000-resample bootstrap 95% CI [.849, .917]). Because the auditor saw the assigned label, these are audit-agreement statistics rather than independent inter-annotator reliability estimates.

The audit covers the major retrieval decisions as well as substantive labels. Agreement was 151/164 (92.1%) for *Non-target*, 159/171 (93.0%) for *Unclassifiable*, 18/18 (100%) for *Multiple*, and 93/107 (86.9%) across the ten substantive labels (Table 5). Corrections were applied to the full counts: 1,931 *Non-target*, 1,388 *Unclassifiable*, 205 *Multiple*, and 1,076 substantive single-category cases. *Non-target* plus *Unclassifiable* account for 72.2% of the sample.

The audiovisual and web-text distributions reflect different modality-specific filters and are compared only within recoverable single-category cases. Vaccination and discomfort are more prominent audiovisually, where precipitating events can be directly visible; sleepiness, hunger, birth, other pain, and overstimulation are proportionally more frequent in web text, where writers often name an attributed cause. Fussiness is comparatively

Initial audit stratum	<i>n</i>	Agree	%
Non-target	164	151	92.1
Unclassifiable	171	159	93.0
Multiple	18	18	100.0
Substantive single category	107	93	86.9
All audited lines	460	421	91.5

Table 5: Human audit of the first 460 enTenTen21 concordance classifications. Agreement is reported after normalization to the declared web-text label inventory.

similar. Sleep-related discourse also exposes a boundary problem: *cry it out* or sleep-training references may describe a caregiving practice without establishing physiological sleepiness as the immediate cause. A post-hoc lexical check confirmed that such discourse spans several labels rather than forming a unique cry type. The comparison therefore concerns contextual recoverability, not population prevalence, and supports keeping *retrieval evidence*, *textual attribution*, and *audiovisual coding* distinct.

4.2 Causative constructions as a precision probe

A complementary query targeted the more explicitly causal construction *MAKE* + *baby/infant/newborn/neonate* + *CRY*. This query was used only as a precision probe and is not included in the 4,600-line distributional counts above. It returned 908 concordance lines. Although the construction should in principle make a causer easier to recover, 274 hits (30.2%) instantiated the recurrent *make Baby Jesus/Jebus cry* formula.¹ Related *Baby X* variants and other figurative or non-causal uses added further noise, and only 145 cases were provisionally mappable to a cry context. Under the contextual scheme used above, such formulaic cases are *Non-target* rather than infant-cry events.

The Baby Jesus/Jebus cases are analytically useful rather than merely accidental noise: they show that lexical proximity and even an apparently causal syntactic frame do not guarantee the intended referent or a literal crying event. Narrower WaC queries therefore still require contextual screening of formulaic, figurative, and pragmatic uses.

4.3 Refining Other Pain

Qualitative inspection of cases grouped under the broad *Other Pain* category during taxonomy development revealed recurrent subcontexts including

¹*Jebus* is the intentional spelling occurring in the recurrent formula, not a typographical error.

teething or oral pain, circumcision and other medical procedures, burns or scalds, and general injury or trauma. These distinctions are not treated as separate top-level labels in the present taxonomy because the available examples remain sparse, but they indicate directions for future refinement of the category.

5 Enriched representation for future interactional analysis

The video-level context labels distinguish eliciting situations that imply different candidate caregiver responses: hunger and sleepiness concern homeostatic regulation; vaccination and other pain provide acute nociceptive contexts; discomfort captures externally elicited non-acute physical discomfort; gas and colic-like crying capture internal gastrointestinal or regulatory contexts; and fussiness or overstimulation concern regulatory or interactional distress. This contextual layer can therefore support later analyses of how caregivers respond to a cry without treating the cry signal itself as a diagnosis. Vaccination work, for example, links caregiver emotional availability and soothing or distress-promoting behavior to infant pain regulation (Atkinson et al., 2015; Badovinac et al., 2018).

For selected future interactional analyses, we propose a source-decoupled representation aligning scene/procedural events, infant vocal activity, visible infant behavior, caregiving action, and minimized or paraphrased adult speech. Cry notation is descriptive rather than diagnostic: @_C/@_c mark relatively louder/softer pulses, _phon/_dys phonated versus dysphonic quality, and _ex/_insp expiratory versus inspiratory phase. Table 7 gives a synthetic example solely to demonstrate the representation; it does not reproduce any individual source interaction.

6 Quality control and audit trail

After duplicate-marked and non-retained rows are excluded, 250 unique classified videos remain. Current validation covers 230/250 (92.0%): 200 retained records have a non-empty explicit audit/validation judgment, including 15 vaccination cases, and the remaining 30 vaccination videos underwent the separate video-level context check described above. For agreement calculation, trivial spelling/spacing differences and parenthetical explanatory notes were normalized. Before adjudication and taxonomy refinement, the initial and

Cry context	Audiovisual		Web text		Δ pp
	<i>n</i>	%	<i>n</i>	%	
Vaccination	45	22.2	10	0.9	-21.2
Discomfort	44	21.7	120	11.2	-10.5
Fussiness	30	14.8	176	16.4	+1.6
Colic-like / Colic	12	5.9	87	8.1	+2.2
Hunger	13	6.4	137	12.7	+6.3
Birth	8	3.9	110	10.2	+6.3
Other Pain	8	3.9	93	8.6	+4.7
Overstimulation	6	3.0	85	7.9	+4.9
Gas	3	1.5	1	0.1	-1.4
Sleepiness	34	16.7	257	23.9	+7.1
Total	203	100.0	1,076	100.0	

Table 6: Distribution of the ten shared contexts in validated audiovisual cases and single-category web-text classifications. Audiovisual *Colic-like* is compared with web-text *Colic*. Percentages are within classified cases, not population prevalence estimates.

Tier	Synthetic representation
PROC	((Injection occurs.))
BABY	@_C_phon_ex → @_C_dys_ex
VIS	((Infant withdraws and raises arms.))
ADULT	((Caregiver moves closer and provides soothing touch.))

Table 7: Synthetic illustration of the proposed source-decoupled representation. No row reproduces an identifiable source interaction.

validation labels agree in 156/230 cases (67.8%; Cohen’s $\kappa = .636$). Restricting the denominator to validation-covered records whose initial label was substantive gives 145/188 agreement (77.1%; $\kappa = .732$).

These values are best interpreted as *audit statistics*, not as reliability estimates for a frozen final codebook. The validation pass was used both to check individual labels and to sharpen category boundaries, and it was not designed as a blinded fresh annotation exercise. Final distributional tables use the validated/adjudicated labels. Of the 230 validation-covered records, 226 receive a current taxonomy label; three no-cry cases and one unresolved *Discomfort/Overstimulation* case do not enter the taxonomy distributions. The two proximity-seeking judgments are nested under *Fussiness*. Exclusion of *Unknown* and *Multiple* therefore leaves 203 single-category audiovisual cases compared with web text. A conventional category-level inter-annotator reliability analysis would require a fresh pass in which multiple coders independently apply the finalized guide.

The disagreement pattern was itself analytically useful: it identified categories whose operational criteria required revision rather than treating all disagreements as interchangeable coding error. The systematic reassignment of initially *Unknown* cases to *Sleepiness*, and the recurrent subtypes observed

within *Fussiness*, are described in Section 3.1.

7 Ethics, data protection, and proposed reuse

The source material consists of publicly accessible videos on YouTube, Instagram, and TikTok. The videos were viewed on their source platforms for research coding; they were not downloaded or retained as audiovisual files in the present phase. Because the recordings involve infants, caregivers, and, in some cases, medical procedures, birth, or visible distress, public accessibility is not treated as sufficient justification for unrestricted reuse or redistribution. The paper therefore reports aggregate results and uses synthetic rather than case-level audiovisual illustrations.

The terms of the three source platforms were checked on 20 September 2026. YouTube restricts downloading, reproduction, distribution, modification, and automated access except where permitted by the service, rights holders, or applicable law.² Instagram prohibits unauthorized access or collection and requires respect for third-party intellectual-property rights.³ TikTok’s EEA terms restrict unapproved automated extraction and make some reuse, including downloading, dependent on platform features and settings.⁴ Public availability is therefore not treated as a blanket licence for local acquisition or redistribution.

The present study therefore keeps the obser-

²YouTube Terms of Service, “Permissions and Restrictions”: <https://www.youtube.com/t/terms>.

³Instagram Terms of Use: <https://help.instagram.com/581066165581870>.

⁴TikTok Terms of Service (EEA/UK/Switzerland): <https://www.tiktok.com/legal/page/eea/terms-of-service/en>.

vational analysis distinct from a possible future preservation phase. Subject to favorable institutional ethical and data-protection review, and only where the applicable platform terms, rights-holder permissions/licences, or legal exceptions permit it, a future phase could temporarily acquire authorized source material solely to create privacy-preserving derivatives. Proposed safeguards include neutral or synthetic facial rendering driven by landmarks/action units, removal or inpainting of identifying backgrounds and visible text, voice de-identification or replacement by transcription, metadata stripping, and re-identification testing. A derivative would not be described as anonymous unless that standard could be demonstrated; otherwise audiovisual derivatives would not be openly released and access would remain restricted.

8 Discussion

The validation results show why taxonomy development and reliability assessment are not separate stages in this type of corpus. The large movement from *Unknown* to *Sleepiness* did not simply add another label; it revealed that the first coding guide had under-specified a visually available evidential domain. Once sleep-state transitions and converging cues such as repeated eye closure or partial arousal were made explicit, a substantial set of previously residual cases became classifiable. Conversely, the refinement of *Fussiness* showed that a useful top-level category can remain internally heterogeneous. Goal-blockage protest, proximity-seeking or separation-related protest, and socially elicited distress can be kept together at the present level while remaining available as analytically meaningful subtypes.

The same validation process suggests a potentially useful cross-cutting distinction between ordinary regulatory frustration and abrupt event discontinuity. Many fussiness cases concern a familiar situation becoming temporarily undesirable or unavailable; overstimulation more often involves sudden, intense, novel, or expectancy-disrupting input. Other pain can share this abruptness while remaining distinguished by an identifiable nociceptive event. Future work can test whether this event-continuity dimension improves annotation reliability or predicts differences in infant or caregiver response before it is incorporated into the coding scheme.

Uploader metadata may also be analytically use-

ful as a separate attribution layer. One audiovisual case remained *Unknown* from the recording itself but was described in uploader metadata as hunger-related. More generally, terms such as *fussy*, *self-soothing*, and *cry it out* can encode a caregiver's interpretation of the infant and a stance toward appropriate response rather than a directly observable cause. This is particularly consequential online, where sleep-training and responsiveness practices are publicly debated and the same crying episode may be framed as sleepiness, protest, a need for proximity, or behavior that should be allowed to resolve without intervention. Rather than using uploader wording to overwrite audiovisual evidence, we therefore treat metadata as an attribution layer that can itself be studied.

9 Limitations

The audiovisual sample is exploratory rather than representative. Candidate links were located through unconstrained, non-random discovery on YouTube, Instagram, and TikTok, so inclusion reflects public posting and platform search/ranking practices. Platform provenance was not retained analytically, preventing platform-stratified counts. No French-language videos occur in the sample, but this does not imply international representativeness. Demographic and clinical information is usually unavailable, recording conditions vary, and source availability changes over time. The distributions therefore describe this web-visible sample rather than population prevalence.

The audiovisual validation procedure also has an important design limitation: it served both as a label audit and as a mechanism for refining the taxonomy, rather than as blinded application of a frozen codebook. The reported values are therefore pre-adjudication audit statistics, not final inter-annotator reliability. Several categories remain small, especially *Gas*, *Birth*, *Other Pain*, and *Overstimulation*. A stronger test would require a fresh independent coding pass using Appendix A, with category-level agreement and a confusion matrix. *Fussiness* is likewise an operational, heterogeneous context label, not a claim that distress is trivial or lacks a cause.

The web-text analysis has different constraints. The 4,600 lines are the first contiguous query block rather than a random sample, and classification uses exported KWIC context rather than full documents. The 10% audit is corrective, not blind,

so 91.5% agreement and $\kappa = .885$ quantify audit agreement rather than independent reliability. The *cry-it-out*/Ferber cases further show that parenting terminology may describe caregiving practice or stance without uniquely identifying the infant's immediate state, particularly for *Sleepiness* and *Fussiness*.

Exact audiovisual reconstruction is intentionally limited by the privacy-preserving design. Source media and identifiable case-level material are not released, and uploader metadata were not systematically retained as an analytic layer. Reproducibility is therefore strongest for the coding scheme, aggregate counts, web-text query, and audit protocol. Future work should retain ethically permissible provenance variables and independently evaluate the frozen taxonomy.

10 Conclusion

We presented an analyzed sample of 250 unique web-derived infant crying videos from YouTube, Instagram, and TikTok, an operational context taxonomy, and a validation audit covering 230 videos (92.0%). Of these, 226 receive current taxonomy labels, and 203 substantive single-category cases enter the cross-modal comparison. The audit provides pre-adjudication agreement statistics while also documenting how the taxonomy was refined: in particular, it exposed the under-specification of sleep-related visual evidence and revealed recurrent regulatory/protest subtypes within *Fussiness*. Reported audiovisual distributions use validated labels only.

A supporting enTenTen21 analysis shows that textual cry context can be recovered at scale through contextual judgment, but also that many web concordances are non-target or leave the eliciting cause unspecified. A human audit of 460 lines (10.0%) covered the full output inventory rather than only substantive cases and yielded 91.5% normalized categorical agreement ($\kappa = .885$). The cross-modal comparison further shows why textual attribution and audiovisual evidence should remain distinct: a cause may be linguistically named in text yet visually unavailable, while an event such as an injection may be directly observable without being verbally identified.

Finally, the synthetic vaccination sequence demonstrates the proposed enriched transcription format without reproducing an individual source interaction. The format keeps contextual labels

distinct from observable infant vocal activity, visible behavior, adult speech, and caregiving action. For Web-as-Corpus research, the study therefore illustrates a workflow in which broad discovery, audiovisual coding, validation-driven taxonomy refinement, contextual web-text classification, multimodal transcription, and release constraints are documented separately.

Acknowledgments

We thank Alex Aguiar, Telmen Altanbagana, Darline Aurivel, Anamika Das, Cassidy Desousa, Alexandra Gelabert, Juliette Guet Dupre, Apolline Huet, Lila Parent, Makizhiny Paul, Danica Penafiora, Charlene Pires, Lison Rose, Clara Silvestre, Niruya Sivanand, Ines Torchon, Paris-Cité undergraduate students for their assistance with locating publicly accessible video links during the initial corpus-discovery stage and for interesting discussions. Their contribution to corpus compilation was limited to identifying candidate links; video coding and analysis were carried out separately by the research team.

This work was supported by the ANR CODIM Project (<https://www.codim-project.org/>) and by the investment program "France 2030" launched by the French Government and implemented by the University Paris Cité as part of its program "Initiative d'excellence" IdEx with the reference (ANR-10-IDEX-0001).

AI use statement

OpenAI ChatGPT was used during manuscript preparation for drafting, restructuring, and language editing. The authors reviewed and revised the resulting text and are responsible for the final content. The separate use of GPT-5.6 Sol for concordance classification is described in Section 4 as part of the study methodology.

References

- Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. In *arXiv preprint arXiv:1609.08675*.
- Nicole H. Atkinson, Hannah Gennis, Nicole M. Racine, and Rebecca R. Pillai Riddell. 2015. *Caregiver emotional availability, caregiver soothing behaviors, and infant pain during immunization*. *Journal of Pediatric Psychology*, 40(10):1105–1114.

- Shaylea Badovinac, Hannah Gennis, Rebecca Pillai Riddell, Hartley Garfield, and Saul Greenberg. 2018. [Understanding the relative contributions of sensitive and insensitive parent behaviors on infant vaccination pain](#). *Children*, 5(6):80.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2006. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.
- Carlo V. Bellieni, Renata Sisto, Duccio M. Cordelli, and Giuseppe Buonocore. 2004. [Cry features reflect pain intensity in term newborns: An alarm threshold](#). *Pediatric Research*, 55(1):142–146.
- Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1547.
- Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 721–725.
- Pamela S. Douglas and Peter S. Hill. 2013. [Behavioral sleep interventions in the first six months of life do not improve outcomes for mothers or infants: A systematic review](#). *Journal of Developmental & Behavioral Pediatrics*, 34(7):497–507.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.
- Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780.
- Alexander M. Goberman and Jason A. Whitfield. 2013. [Acoustics of infant pain cries: Fundamental frequency as a measure of arousal](#). *Perspectives on Speech Science and Orofacial Disorders*, 23(1):18–26.
- Michael Gradisar, Kate Jackson, Nicola J. Spurrier, Joyce Gibson, Justine Whitham, Anne Sved Williams, Robyn Dolby, and David J. Kennaway. 2016. [Behavioral interventions for infant sleep problems: A randomized controlled trial](#). *Pediatrics*, 137(6):e20151486.
- Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056.
- Katherine L. Guyon-Harris, Kathryn L. Humphreys, and Charles H. Zeanah. 2021. [Adverse caregiving in early life: The trauma and deprivation distinction in young children](#). *Infant Mental Health Journal*, 42(1):87–95.
- Mohammed Hammoud, Melaku N. Getahun, Anna Baldycheva, and Andrey Somov. 2024. [Machine learning-based infant crying interpretation](#). *Frontiers in Artificial Intelligence*, 7:1337356.
- Chunyan Ji, Thosini Bamunu Mudiyanse, Yutong Gao, and Yi Pan. 2021. [A review of infant cry analysis and classification](#). *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):8.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. [The sketch engine: Ten years on](#). *Lexicography*, 1:7–36.
- Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. In *Computational Linguistics*, volume 29, pages 333–347.
- Ana Laguna, Sandra Pusil, Àngel Bazán, Jonathan Adrián Zegarra-Valdivia, Anna Lucia Paltrinieri, Paolo Piras, Clàudia Palomares i Perera, Alexandra Pardos Véglia, Oscar Garcia-Algar, and Silvia Orlandi. 2023. [Multi-modal analysis of infant cry types characterization: Acoustics, body language and brain signals](#). *Computers in Biology and Medicine*, 167:107626.
- Marguerite Lockhart-Bouron, Andrey Anikin, Katarzyna Pisanski, Siloé Corvin, Clément Cornec, Léo Papet, Florence Levréro, Camille Fauchon, Hugues Patural, David Reby, and Nicolas Mathevon. 2023. [Infant cries convey both stable and dynamic information about age and identity](#). *Communications Psychology*, 1(1):26.
- Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech*, pages 2616–2620.

- Joanna J. Parga, Sharon Lewin, Juanita Lewis, Diana Montoya-Williams, Abeer Alwan, Brianna Shaul, Carol Han, Susan Y. Bookheimer, Sherry Eyer, Mirella Dapretto, Lonnie Zeltzer, Lauren Dunlap, Usha Nookala, Daniel Sun, Bianca H. Dang, and Ariana E. Anderson. 2020. [Defining and distinguishing infant behavioral states using acoustic cry analysis: Is colic painful?](#) *Pediatric Research*, 87(3):576–580.
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.
- Anna M. H. Price, Melissa Wake, Obioha C. Ukoumunne, and Harriet Hiscock. 2012. [Five-year follow-up of harms and benefits of behavioral infant sleep intervention: Randomized trial.](#) *Pediatrics*, 130(4):643–651.
- Sijmen A. Reijneveld, Marcel F. van der Wal, Emily Brugman, Remy A. Hira Sing, and S. Pauline Verloove-Vanhorick. 2004. [Infant crying and abuse.](#) *The Lancet*, 364(9442):1340–1342.
- Xuewen Yao, Megan Micheletti, Mckensey Johnson, Edison Thomaz, and Kaya de Barbaro. 2020. Infant crying detection in real-world environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3).
- Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. In *arXiv preprint arXiv:1606.06259*.
- AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246.

A Cry-context coding guide

The labels classify the *eliciting context of crying*, not an acoustic cry type or clinical diagnosis. Evidence is ranked as follows: (1) participant attribution or in-video text; (2) a directly observable proximal event or condition; (3) converging infant behavior; and (4) cry characteristics as supporting evidence only. Uploader titles, descriptions, tags, and search terms are discovery metadata and do not determine the audiovisual label. The scheme is informed by work distinguishing acute pain, hunger, fussiness, discomfort, colic, and related contexts (Bellieni et al., 2004; Goberman and Whitfield, 2013; Parga et al., 2020; Laguna et al., 2023; Hammoud et al., 2024; Ji et al., 2021). The present taxonomy additionally separates GAS from externally

elicited DISCOMFORT, since gastrointestinal gas or straining need not be painful and is not necessarily caused by an external event.

B Web-text contextual classification protocol

For each enTenTen21 line, only the exported *Left + KWIC + Right* context was considered. The classifier was instructed to make a contextual judgment rather than match keywords: first determine whether the passage describes an actual infant-crying event; assign NON-TARGET to figurative, fictional, quoted, imagined, device-related, lyric, or otherwise non-event uses; assign UNCLASSIFIABLE when an actual cry is present but its eliciting context cannot be recovered; use MULTIPLE only when more than one context is independently supported; otherwise select one of the ten shared cry-context labels. Causes were not inferred from setting or co-occurrence alone (e.g., feeding did not by itself license HUNGER, and a distressing scene did not by itself license OTHER PAIN). Each decision returned a label, High/Medium/Low confidence, and a short contextual justification.

Label	Assign when	Do not assign from
Vaccination	A visible injection/vaccination is temporally linked to crying, or participant speech/in-video text explicitly identifies it.	Clinic setting or a needle elsewhere in the scene alone.
Other Pain	A painful procedure, injury, or clearly painful pressure/compression is visibly or explicitly established and linked to the cry (e.g. heel prick, blood draw, injury).	Intensity, roughness, ordinary handling, or presumed pain without an identifiable painful event.
Birth	The cry is a spontaneous first/early neonatal cry in the immediate post-delivery transition. Routine cleaning may coexist when it is not the elicitor.	If active chest pounding, hard rocking, or another intervention is used to elicit a delayed cry, code the immediate elicitor as DISCOMFORT or OVERSTIMULATION. An unresponsive infant with no cry is not a crying case.
Hunger	Participant speech/in-video text attributes hunger, or a clear pre-feeding sequence plus converging rooting/searching/sucking behavior supports it; feeding that promptly resolves the episode may corroborate.	A bottle/breast/feed somewhere in the video, post-feeding crying, or uploader metadata alone.
Sleepiness	Tiredness/sleep is explicitly attributed or a clear sleep transition, partial-arousal state, or settling sequence is supported by drowsiness, yawning, eye-rubbing, repeated eye closure, etc.	Night-time, crib/bedtime, eventual sleep, or a sleep-training/cry-it-out frame alone.
Gas	Gas, trapped air, difficulty burping, reflux, constipation/straining, or a related gastrointestinal process is explicitly attributed or strongly established, with converging behavior such as leg flexion, straining, regurgitation, or attempts to burp or pass gas. Pain is not required.	Burping, feeding, passing gas, or tummy touching alone; do not equate gas with colic-like crying or with externally elicited discomfort.
Discomfort	An externally caused, non-acute physical condition or handling is clearly implicated: wet/soiled diaper, bathing/cleaning, uncomfortable position or clothing, temperature exposure, skin irritation, restrictive mittens, restraint, or a sustained difficult posture.	Internal gastrointestinal states such as gas, reflux, or straining; or diaper change, bathing, clothing, mittens, handling, or posture merely being present without evidence that the external condition contributes to crying.
Overstimulation	A sudden/intense sensory or social stimulus immediately precedes the cry; startle, withdrawal, turning away, escalation with continued stimulation, or relief when it stops may support the label.	Background television/music, ordinary conversation, people in the room, or ambient noise without a temporal relation.
Fussiness	Regulatory/protest distress after specific homeostatic, gastrointestinal, externally elicited discomfort, and nociceptive causes are unsupported; subtypes may include goal blockage, proximity-seeking, and socially elicited distress.	Not an unexplained-cry default. The label does not imply triviality or endorse non-response; do not assign from sleep-training framing or caregiver non-response alone.
Colic-like	Prefer explicit participant/in-video attribution to colic. Without it, require sustained or recurrent difficult-to-console crying in a sufficiently extended recording, with other specific contexts excluded; otherwise cue-free clips under 30 s are UNKNOWN.	Cry acoustics, intensity, apparent gas, or unexplained crying alone. The label is descriptive, not a clinical diagnosis.
Multiple	Two or more different contexts independently satisfy their own criteria within the same item.	Uncertainty between alternatives; ambiguous alternatives are UNKNOWN, not MULTIPLE.
Unknown	No category meets its evidential criterion, or cues conflict so that the eliciting context cannot be assigned confidently.	A “most likely” guess made only to avoid an unknown label.

Table 8: Concise operational guide for audiovisual cry-context coding.

C Sample of the web-text classification file

Table 9 reproduces the first 100 rows of the 4,600-line enTenTen21 classification file with all exported fields. The *Left* and *Right* columns are reproduced in full rather than abridged. A blank *Validation* cell means that the human audit accepted the assigned classification unchanged.

Table 9: First 100 rows of the 4,600-line enTenTen21 contextual-classification file.

Line	Reference	Left	KWIC	Right	Classification	Conf.	Contextual basis	Validation
1	chs.ca	. </s><< However, additional transmitters can be added to receive alerts for the telephone/TTY ringing (BEI1023).	baby crying	Stent Call Meditation Series </s><< What it is: Smoke detector with society's lies on Capitalism. </s><< Why don't these winners get on	Non-target	High	Device/accessibility example: no infant crying event is being described.	
2	hnn.us	pro-active. </s><< This country is like living with a bunch of adult	babies crying	but the mother knows to stay silent. </s><< If this camp, fewer than	Non-target	High	Metaphorical use: adults compared to crying babies.	
3	hnn.us	this brutal system. </s><< A mother with a young baby falls in the	baby cries	and all prisoners will survive. </s><< Even the commandant	Other Pain	High	The baby and mother have just been beaten by a	
4	cme.ie	lifting feet and both are beaten mercilessly by a guard. </s><< The	babies crying	half overboard, work their way through like	Non-target	High	Context: list of sounds in an artistic work; no infant cause/cause.	
5	nhs.uk	Aljorfi, Daihiel Shihama. </s><< WAC's hosting of this group does	baby cries	a lot but there's, no obvious cause. </s><< It's a common problem	Colic	High	Explicitly defines colic as prolonged crying without obvious cause.	
6	nhs.uk	problem that should get better on its own. </s><< Colic is when a	babies cry	GP if		High	Context is a colic diagnostic description.	
7	nhs.uk	or see a GP if you're worried. </s><< Check if your baby has colic		days a week for at least 1 week. </s><< Immediate action	Colic	High	Abnormal cry is flagged medically, but no cause is given.	Medically concerning cry (abnormal cry)
8	hse.ie	hours a day </s><< Colic can be very upsetting for parents. </s><<	baby crying	so intensely, and it can seem like your baby is in pain. </s><< Signs	Colic	High	Explicit signs/symptoms of colic.	
9	u2.com	old. </s><< A baby has colic: if they cry: </s><< For more than 3	baby cries	has	Non-target	High	Song lyric rather than a described infant-crying episode.	
10	u2.com	the world but you know better And that all we have is immortality	baby cries	You argue because you can't accept Love is all we have	Non-target	High	Song lyric rather than a described infant-crying episode.	
11	hse.ie	Sure I close your eyes Love and love is all we have left A	baby crying	The only thing that can be kept Love is all we have left </s>	Non-target	High	Bereaved person thinks they hear the baby; no actual crying event.	
12	omn.org	time not to be alive Love and love is all we have left A	baby cried	hercane. </s><< They will give information to help you to make	Unclassifiable	High	An actual baby cries, but the cause is not stated.	
13	tuq.ca	that you experience can include disbelief, anger, guilt and grief.	baby's cry	, the phone ring, everything happened at once. </s><< But once it	Unclassifiable	Medium	Cry is intense and hunger/colic are ruled out, but no alternative cause is supplied.	
14	rt.com	Re. </s><< I want to put my prayers for the people of the		could leak out. </s><< Her head is flung back in a motion,	Non-target	High	Parental sound effect used rhetorically.	
15	cbc.ca	prawns and looks up at Penn with big, mournful eyes. </s><< The	babies cried	than a military grant who thinks for him or herself. </s><< They're	Other Pain	High	Babies are crying while exposed to tear gas.	
16	cbc.ca	themselves! </s><< (Glass shatters. </s><< Woman screams.	baby crying	, tears streaming from the gas, a Reuters witness said. </s><< "They	Non-target	High	Dream/imagined baby crying.	
17	kk.org	</s><< Protesters were caught between the Mexican and U.S. authorities,	Babies cry	with the fourth. </s><< When a kid is sick and can't breathe well	Multiple	High	Explicitly lists hunger, tiredness, diaper and gas as possible causes.	
18	iml.ie	they were babies, these things were super helpful. </s><<	babies crying	weren't as innocent as you think? </s><< New research, published in	Non-target	Medium	Rhetorical/media example rather than a specific infant-crying context.	
19	iml.ie	from the Priory Hall residents. </s><< The IMO should be pushing	baby's cries	the journal Evolution, Medicine, and Public Health,	Unclassifiable	Medium	Discussion of infant crying research, but no cause for the cry is given.	
20	dw.com	and nourish your offspring... and this is all the consolation you need	baby cry	, your besamnik may appear at the ready ... </s><< ... even if it's not	Unclassifiable	High	Real baby crying is invoked, but only as a trigger for milk let-down; cause unknown.	
21	dw.com	don't breathe air in the womb and the cycle starts all over again,	baby crying	from underneath her. </s><< When he turned over the corpse, he	Unclassifiable	Medium	Infant is found under a corpse in disaster debris; distressed, but the immediate cause of crying is not established.	
22	iml.ie	</s><< If you hear a	babies crying	</s><< Said Dr Sheridan Pereira, but parents could end up depressed. b4.	Multiple	Medium	The passage explicitly frames newborn crying as having multiple aetiologies.	
23	bgrrn	important, she said. </s><< Maternal depression was not the cause of	baby crying	user's smartphone. </s><< Users can listen to the	Non-target	High	AI sound-detection example.	
24	nj.gov	thanks to an improved AI. </s><< So, if the device detects a	Baby Cries	(Shaken Baby Syndrome) </s><< When a Baby Cries (Shaken Baby	Unclassifiable	Medium	Campaign concerns coping with a crying baby, but no cause of crying is specified.	
25	nj.gov	Science in TIMSS 2015 </s><< Part of </s><< Share </s><<	Baby Cries	Syndrome) - Campaign warns parents/caregivers to never shake a	Unclassifiable	Medium	Campaign concerns coping with a crying baby, but no cause of crying is specified.	
26	ltd.org	These materials can be downloaded or ordered through DCF	baby cry	. </s><< On On Pipes Space City RA (Houston) Hamersley Hash	Non-target	Low	No clear infant context; likely figurative/opaque use in a hash-related passage.	

Continued on next page

Table 9 – continued from previous page

Line	Reference	Left	KWIC	Right	Classification	Conf.	Contextual basis	Validation
60	ndt.com	County Sheriff's Office deputies late June 6. </s></s> They responded baby crying after receiving a 911 call from someone who heard the	baby crying	</s></s> Forsyth Sheriff Ron Freeman said at the time that the discovery of the baby was a miracle. </s></s> "It was, without doubt,	Unclassifiable	Medium	Abandoned baby is heard crying; cause is not stated.	Abandoned baby
61	ndt.com	</s></s> About five minutes after the call, the caller told emergency services that the baby was breathing. </s></s> "That's the	baby crying	</s></s> We're getting vitals from mom right now," the caller said.	Birth	Medium	Emergency context strongly suggests a just-delivered baby; crying confirms breathing.	Emergency context strongly suggests a just-delivered baby; crying confirms breathing.
62	esf.edu	</s></s> Hearing Ear Alerts persons with hearing impairments to specific sounds within the environment, such as a	baby crying	</s></s> Phoenix police have launched a sexual assault	Non-target	High	Service-animal/accessibility example.	Service-animal/accessibility example.
63	apl.org	she is on a collision course to meet them. </s></s> Twenty-five years ago, police were called to 16 Cleyne Walk with reports of a	baby crying	</s></s> A telephone ringing, a fire alarm, or door bell. </s></s> Fetch & Carry.	Unclassifiable	High	Healthy infant is heard crying in crib; no cause supplied.	Healthy infant is heard crying in crib; no cause supplied.
64	edj.net	</s></s> My family and in-laws came to the house here during that time. babies cried	babies cried	</s></s> Assistants mobility impaired persons by obtaining and happily cooing in her crib in the bedroom. </s></s> Downstairs in the	Unclassifiable	High	Babies cry during a frightening/disaster situation, but cause is not specific enough.	Babies cry during a frightening/disaster situation, but cause is not specific enough.
65	bprc.org	The there are fireworks. </s></s> It plays me fan sounds through my wife's baby cries	baby's cries	</s></s> Don Rafael's house was indeed small. </s></s> Made of brick, adobe	Unclassifiable	Medium	Comedic song/quiz line about a white-noise machine.	Comedic song/quiz line about a white-noise machine.
66	ch.org	When my defenceless, even to the point of lacking that minimal form of defence baby's cries consisting in the poignant power of a newborn	baby's cries	</s></s> I figure. </s></s> MORENO: That is a white noise machine. </s></s> COULTON: White noise machine - that is correct. </s></s> EISENBERG: I do find it	Non-target	High	Rhetorical comparison involving newborn cries, not a crying episode.	Rhetorical comparison involving newborn cries, not a crying episode.
67	icld.us	</s></s> your baby's fever is higher than 10/26 when taken in the	baby cries	</s></s> or screams for 3 or more hours after the DTaP shot and can't be	Vaccination	High	Prolonged inconsolable crying is explicitly post-DTaP shot.	Prolonged inconsolable crying is explicitly post-DTaP shot.
68	chr.com	Black Panther's Alpha Flight #106 starred with the team fighting Mr. baby crying	baby crying	</s></s> from inside of a dumpster. </s></s> Putting the newborn out, he	Non-target	High	Comic-book fictional scene.	Comic-book fictional scene.
69	uhcg.us	</s></s> Even the jacksals feed their young, but not my people. Israel, babies' cries	babies' cries	</s></s> The children's tongues stick to the roofs of their mouths for	Hunger	High	The surrounding passage explicitly describes thirst/deprivation.	The surrounding passage explicitly describes thirst/deprivation.
70	uhcg.us	</s></s> The children's tongues stick to the roofs of their babies cry	Babies cry	</s></s> for bread, but no one can give them any. </s></s> Those who used to	Hunger	High	Explicit: babies cry for bread.	Explicit: babies cry for bread.
71	jwa.org	months for thirst, for there is not A drop of water left. </s></s> other way. </s></s> These images are impossible to make real without	baby's cries	</s></s> eat fastidiously are begging in the streets for anything at all. </s></s> become overly disruptive or to help set up kiddush when the	Unclassifiable	High	Realistic generic crying in synagogue; cause unknown.	Realistic generic crying in synagogue; cause unknown.
72	zcp.org	in the age of 2 to 4 months. </s></s> The infants die during sleep and	baby cried	</s></s> when it happened. </s></s> There is no treatment or even a clear	Non-target	High	SIDS passage explicitly says there may have been no crying.	SIDS passage explicitly says there may have been no crying.
73	lapm.ca	if the Germans were close by. </s></s> His father disagreed, and felt	baby's cries	</s></s> specific prevention of SIDS, even with years of medical research	Unclassifiable	High	Baby cries while people hide from soldiers; cause of cry is not stated.	Baby cries while people hide from soldiers; cause of cry is not stated.
74	stt.edu	these options doesn't guarantee that the video conferencing tools	baby cries	</s></s> put aboard trains. </s></s> While being transported,	Non-target	High	Hypothetical advertising/privacy example.	Hypothetical advertising/privacy example.
75	chr.com	aren't collecting data on your surroundings. </s></s> If a	baby crying	</s></s> your pet pops up on screen, ads for pet supplies soon. </s></s> If	Non-target	High	Comic-book imagery.	Comic-book imagery.
76	bprc.org	massive face taking the majority of the focus away from a	baby crying	</s></s> and a sky scraper crumbling to the ground. </s></s> Pylocke is	Non-target	High	Generic background-noise example in remote meetings.	Generic background-noise example in remote meetings.
77	bprc.org	says quarantine quite like hearing a stray tune signal or a	babies crying	</s></s> in the background of your work meeting or virtual family gathering.	Non-target	High	Babies are crying in a crowded/cold setting, but the line does not establish why.	Babies are crying in a crowded/cold setting, but the line does not establish why.
78	jnj.com	Nights are the worst. </s></s> BRITT: You have a bunch of people	babies crying	</s></s> </s></s> Comic Nore Davis has heard it all because when New	Unclassifiable	Medium	Generic background-noise example in remote meetings.	Generic background-noise example in remote meetings.
79	bprc.org	hand either to go to sleep or wake up. </s></s> Could these social facts	baby's cries	</s></s> a quilt on my bed, and I got a real thick housecoat. </s></s> And I	Unclassifiable	High	Abandoned baby is found after cries are heard; precise cause is not stated.	Abandoned baby is found after cries are heard; precise cause is not stated.
80	life.ca	be connected? </s></s> I strongly suspect they are." </s></s> ~ John	baby's cry	</s></s> is precisely as serious as it sounds." </s></s> ~ Jan Liedloff </s></s> Multiple	Multiple	Medium	The adjacent quotation lists warmth, milk and security as newborn needs associated with crying.	The adjacent quotation lists warmth, milk and security as newborn needs associated with crying.
81	chr.com	him that their connection will transcend the generations, however, and	baby's cry	</s></s> ... </s></s> 1 Episode 39: "Venom of the Red Lotus" (9/6) </s></s> This is the highest-rated The Legend of Korra episode of all, and it's	Non-target	High	Television episode sound cue.	Television episode sound cue.
82	wvi.org	how I tried to give the baby water to keep her calm. </s></s> But I	baby crying	</s></s> and was on the verge of giving up the notion of exclusive breastfeeding	Unclassifiable	Medium	The mother considers giving water because the baby is crying, but the text does not establish hunger/thirst as the cause.	The mother considers giving water because the baby is crying, but the text does not establish hunger/thirst as the cause.
83	pets.ca	hers. </s></s> Even cats that have a high percentage of siamese blood	baby crying	</s></s> when a member of the "Mamanetvor" (Grandmother)	Unclassifiable	High	Cat vocalization compared with a baby crying.	Cat vocalization compared with a baby crying.
84	pets.ca	attention, and for some cats they just like to talk. </s></s> They sound like a	baby crying	</s></s> , and this will go on and on if you're not giving them enough attention. Non-target	Non-target	High	Cat vocalization compared with a baby crying.	Cat vocalization compared with a baby crying.
85	pets.ca	some people understand what I mean by sounds like a	baby crying	</s></s> ... </s></s> I mean sounds like a baby crying... just as loud just as	Non-target	High	Cat vocalization compared with a baby crying.	Cat vocalization compared with a baby crying.
86	npaf.ca	mean by sounds like a baby crying... </s></s> I mean sounds like a	babies cry	</s></s> ... just as loud just as annoying if you can't handle it. </s></s> They	Non-target	High	Generic statement that crying communicates infant needs; no single cause.	Generic statement that crying communicates infant needs; no single cause.
87	jijg.net	in this one because I'm packing a lot of different factors up into this	babies crying	</s></s> : It's the best (and only) way for them to communicate their needs at	Unclassifiable	High	Generic list of distracting background noises.	Generic list of distracting background noises.
88	voy.com	doctor said "the best last thing we heard was a	the baby's cry	</s></s> this tender age. </s></s> And as parents, we're biologically	Non-target	High	Baby cries immediately after the mother pushes during delivery.	Baby cries immediately after the mother pushes during delivery.
89	asac.us	mothers with view pictures of their baby and hear their	the baby's cry	</s></s> Then they started to get the baby's needs. </s></s> "It's the best (and only) way for them to communicate their needs at	Non-target	High	Recorded baby cries are experimental stimuli.	Recorded baby cries are experimental stimuli.
90	voy.com	8:45pm In reply to: Nikki "s message, "letting	baby cry	</s></s> ? and other questions." </s></s> There are a total of four visits over a	Unclassifiable	Medium	Forum thread concerns letting a baby cry, but this concordance line does not establish the cause.	Forum thread concerns letting a baby cry, but this concordance line does not establish the cause.
91	voy.com	the advice given was excellent and the little bit your	baby cries	</s></s> is perfectly normal. </s></s> Your "baby's first" with Santa is only the	Unclassifiable	High	Normal baby crying is mentioned, but no cause is identified.	Normal baby crying is mentioned, but no cause is identified.
92	voy.com	5:42pm In reply to: Nikki "s message, "letting	baby cry	</s></s> ? and other questions..." </s></s> on 12/6/06 12:18pm I work in the	Unclassifiable	Medium	Forum thread concerns letting a baby cry, but this concordance line alone does not establish a sleep-related cause.	Forum thread concerns letting a baby cry, but this concordance line alone does not establish a sleep-related cause.

Continued on next page

Table 9 – continued from previous page

Line	Reference	Left	KWIC	Right	Classification	Conf.	Contextual basis	Validation
93	voy.com	Thread Next Message Previous Message Date Posted: 12/11/06 6:31pm In reply to: Nikki's message, "letting them cry"	baby cry	? and other questions. " </s></s> on 12/6/06 12:18pm Hi. </s></s> I think Luke", advice about sleeping is pretty sound (or not intended </s></s> They are unable to look at a pregnant woman the same Non-target anyone. </s></s> These are not threats or ways of trying to get you to think	Sheepiness	High	Reply explicitly discusses sleeping in the "letting baby cry" thread.	
94	voy.com	</s></s> They wake up in the middle of the night convinced they heard a	baby crying	? and other questions. " </s></s> on 12/6/06 12:18pm Are you breast feeding or bottle feeding? </s></s> If breast feeding you might find the healthy mother/father responds with empathy. </s></s> They literally feel distress - but not for themselves rather for the " written by Lennon and McCartney. </s></s> Mat. a reckless drifter pushing 40, shows up unexpectedly at his gainfully employed ; and they can't reach them. </s></s> Stuff like that. </s></s> Abortion Non-target has a huge emotional impact on a woman. </s></s> You may be able to make	Non-target	High	Imagined cry in a description of post-abortion distress.	
95	voy.com	Previous Thread Next Message Previous Message Date Posted: 12/6/06 9:44pm In reply to: Nikki's message, "letting	baby cry	of another it may the real evocation of a similar feeling in themselves in response to another. </s></s> (Empathy). </s></s> When the Bandana. </s></s> Released in the US by Sire on CD (40135-2) in a digi-pak featuring a blow-up of the artwork of BAD1015. </s></s> "Cry	Hunger	Medium	Reply to crying thread immediately asks about breast-feeding versus bottle-feeding, making feeding/hunger the salient hypothesis.	Unknown
96	bgmi.us	a bunch of men trying to gang-rape me, but at the last minute, I got away. </s></s> Women have told me of nightmares where they hear	baby cries	reminds you of yet another need, and your young toddlers seek your affection and time for play. </s></s> On top of that, your	Unclassifiable	High	Generic empathy example; cause of infant crying unspecified.	
97	evo.org	Bandana. </s></s> Released in the US by Sire on CD (40135-2) in a digi-pak featuring a blow-up of the artwork of BAD1015. </s></s> "Cry	Baby Cry	terribly when they're held wrong, because some instinct says, "This jerk is going to drop me, or "My neck is going to	Non-target	High	Song title/lyric reference.	
98	voy.com	away. </s></s> Women have told me of nightmares where they hear	babies crying	;	Non-target	High	Nightmare about hearing babies crying.	
99	app.com	? </s></s> Last night's dishes, last week's laundry, and yesterday's schoolwork all clamor for your attention. </s></s> Your	baby's cry	;	Unclassifiable	High	Actual household baby cry, but the need causing it is not specified.	
100	tcj.com	his was a mortal enemy. </s></s> And I really do believe that infants have just whom to trust. </s></s> Some instinct tells them. </s></s> We know	babies cry	;	Discomfort	High	Crying is attributed to being held incorrectly/feeling physically insecure.	