NAIST LIFE STORY: A Seven-Year Crowdsourced Dataset of Japanese Emotion-related Episodes

Kazuhiro Ito, Junko Hayashi, Hiroyuki Nagai, Shoko Wakamiya, Eiji Aramaki
Nara Institute of Science and Technology (NAIST)
Ikoma, Nara, Japan
{ito.kazuhiro.ih4, hayashi.junko.hh5}@is.naist.jp,
hiro.nagai@naist.ac.jp, {wakamiya, aramaki}@is.naist.jp

Abstract

Existing emotion datasets have supported a wide range of NLP tasks, but most are static resources that capture language use only at the time of their creation. As a result, they cannot represent how emotional meanings shift in response to cultural and social change. To address this limitation, we present NAIST LIFE STORY, a seven-year collection of Japanese emotion-related episodes that reflect contemporary topics across multiple years. Since 2017, 1,000 crowdsourced participants per quarter have written short texts describing personal experiences associated with seven emotions: anger, anxiety, disgust, trust, joy, sadness, and surprise. The dataset currently spans 28 periods and includes gender and age information for each participant. Analyses reveal systematic differences in text length and lexical diversity across emotions, as well as clear temporal trends linked to major events such as the COVID-19 pandemic. A preliminary experiment with a large language model shows that using this dataset as contextual evidence improves time-aware emotion inference, demonstrating its value for studying the evolving relationship between emotion and language.

Keywords: Emotion, Language resources, Japanese, Crowdsourcing, Diachronic semantics

1. Introduction

Emotion analysis has become a core component of many NLP tasks, including fine-grained emotion classification (Demszky et al., 2020), affect intensity estimation (Mohammad et al., 2018), emotion recognition in conversations (Li et al., 2017; Poria et al., 2019), and identifying triggers of emotion (Xia and Ding, 2019). Accordingly, a wide variety of emotion-related datasets have been developed to support these tasks (Plaza-del Arco et al., 2024; Koufakou and Nieves, 2025).

Recent research further shows that the emotional meaning of words is not static but changes over time. Hamilton et al. (2016) examined word polarity over 150 years and found that more than five percent of non-neutral words reversed polarity, shifting from positive to negative or vice versa. For instance, “awful” once conveyed a sense of awe but now typically means “terrible” or “very bad.” Other work has traced semantic shifts in emotion terms such as “anger” and “happiness” across a century (Xu et al., 2021), showing that even so-called basic emotions like “anger” and “fear” are not immune to change, while context-dependent terms such as “hysteria” or “exhilaration” shift especially strongly.

Despite these findings, most emotion lexicons and labeled datasets remain static resources that reflect the language of their time of creation. As a result, they cannot adequately capture long-term semantic change or shifts linked to specific cultural and social events (Hamilton et al., 2016). Such shifts also reduce model robustness, as models trained on one period often degrade when applied to another (Jaidka et al., 2018). This limitation persists even in the era of large language models (LLMs): although LLMs encode broad knowledge, detecting diachronic semantic shifts still requires explicit longitudinal evidence (Periti et al., 2024). In particular, it remains difficult to capture the emergence and consolidation of new words tied to major social changes, such as those related to the COVID-19 pandemic, and to track how their emotional connotations evolve (Davis, 2023). Addressing this challenge calls for mechanisms that

Figure 1: An overview of the dataset collected through crowdsourcing and its applications. The upper panel presents example episodes for *Joy* and *Disgust* across years. The lower panel shows temporal trends of their key terms, illustrating how emotional expression changes with major social events.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6573–6584
11–16 May 2026.

can reflect and update emotional meanings in a timely and context-sensitive manner.

To this end, we present the NAIST LIFE STORY, a quarterly collection of emotion-elicited episodes from 1,000 participants per round that reflects contemporary topics in Japan (Figure 1). The dataset covers seven core emotions (*anger*, *anxiety*, *disgust*, *trust*, *joy*, *sadness*, and *surprise*) and captures diverse personal experiences associated with each. By providing repeated collections over multiple periods, the dataset enables analyses that consider temporal changes in emotion expression.

2. Related Work

A large body of work provides resources for emotion analysis, from psychological dictionaries to emotion-labeled datasets. LIWC (Pennebaker et al., 2001) (resource: Pennebaker et al. (2015)) offers counts for psychologically meaningful categories and has been extensively used across psychology and social science, yet its dictionary is defined at a point in time and does not track semantic drift by design (Tausczik and Pennebaker, 2010). The NRC Emotion Lexicon (Mohammad and Turney, 2013) (resource: Mohammad and Turney (2013)) supplies word to emotion associations created by crowdsourcing and has become a standard lexical resource in NLP. Within Japanese, WRIME annotates social media posts with intensities for eight emotions from both writer and reader perspectives and provides a valuable benchmark for modeling emotion at a single time slice (Kajiwara et al., 2021) (resource: Kajiwara et al. (2021)). These resources are widely used, but they are typically static snapshots and therefore do not support analysis of changes in the coupling between words and emotions.

Episode style resources exist across communities but are typically not designed for frequent, short-interval collection. The British Library's National Life Stories (British Library, n.d.) preserves personal histories through thematic interview programmes, recording in-depth life episodes from professionals and ordinary individuals across various fields. The Pangloss Collection (Michailovsky et al., 2014) (resource: Michailovsky et al. (2014)) documents languages via oral texts with transcripts and translations, capturing traditional stories and everyday speech from native speakers of diverse, often endangered languages. Large participatory projects such as Voices of the Poor (Narayan et al., 2000) also compiled life stories on poverty across countries, but as one-off or multi-year efforts. Closer to our aims, the CoSoWELL corpus (Kyröläinen et al., 2023) contains written life stories by older adults across several sessions before and during COVID-19, with demographic and psychosocial measures. It shows the value of episodes for studying well-being, but its design focuses on a specific cohort and period rather than ongoing short-interval collection. Crowdsourced episode emotion corpora also exist, such as Troiano et al. (2019, 2022) (resource: Troiano et al. (2019)), who collected short event descriptions for target emotions with validation by independent raters. These resources link events and emotions effectively, yet they were released in single waves rather than repeated collections.

In summary, existing lexicons and corpora support many forms of emotion analysis, but they are generally static or collected as one-off studies, which limits their ability to detect short-term, population-level changes in emotion word associations. Episode style resources show the power of life story for capturing psychological and social phenomena, yet most are not organized as recurring collections with fixed and short intervals. To our knowledge, there is no resource that periodically gathers emotion elicited life stories with new participants each round so as to reflect contemporary topics and enable analysis of shifts over successive periods. Our dataset fills this gap.

3. NAIST LIFE STORY dataset

3.1. Data Collection

The data collected here are released as a NAIST LIFE STORY [1]. Examples of the data are shown in Table 1. All episodes in the dataset were originally written in Japanese. All examples shown in this paper are translated into English for readability.

Using Yahoo! Crowdsourcing[2], we collected free-response descriptions of emotional episodes from a wide range of participants, allowing us to capture natural expressions, including slang and recent trends, across diverse age and gender groups. From 2017 to 2024, data were collected four times a year, generally in February (Season 1), May (Season 2), August (Season 3), and November (Season 4). The actual collection months varied slightly across years. In 2017, only Season 4 was conducted, and in 2024, at the time of writing, only Seasons 1-3 had been completed, resulting in 28 rounds in total. Each round recruits 1,000 participants.

We collected three types of information: **(1) episodes** recalled from emotions, **(2) gender**, and **(3) age**.

(1) Episodes. Participants were asked to write about an episode associated with specific emotion words presented to them. In this study, an episode

[1]Available at https://sociocom.naist.jp/life-story-data/

[2]https://crowdsourcing.yahoo.co.jp/

Emotion	Gender	Age	Episode (example)
Anger	Male	50s	It looks like the event I was planning to join might get canceled because of the coronavirus.
Anxiety	Female	20s	I'm worried because my company's not doing well, and I'm not sure if I should keep working there.
Disgust	Female	60s	I know I'm not perfect myself, but I can't stand how some old people act so rudely and selfishly.
Trust	Male	30s	I feel I can really trust people who give me good advice when I'm struggling with my future.
Joy	Male	40s	I thought staying home with my cat would be boring, but it turned out to be surprisingly fun. I found lots of ways to play together.
Sadness	Female	40s	I spent over a year dealing with surgeries and hospital stays, and just when I thought it was finally over, I ended up with a fracture.
Surprise	Female	10s	I went shopping at the mall for the first time in a while, and the whole interior had completely changed.

Table 1: Examples of data collected in our dataset (originally in Japanese, translated into English). Each entry includes emotion, demographic information, and the recalled episode. Episodes are translated into English for illustration. All examples are from 2020 Season 1.

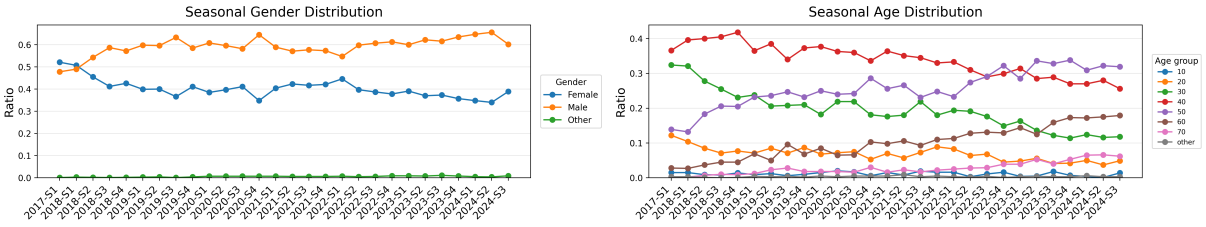

Figure 2: Participant statistics. Left: gender ratios by year. Right: age distribution by year.

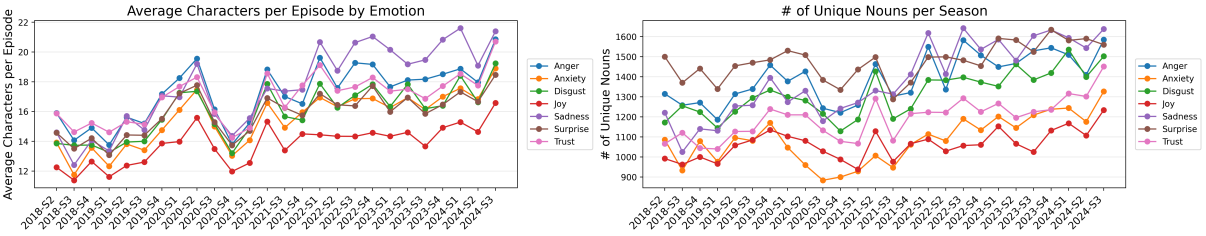

Figure 3: Dataset statistics. Left: average characters per episode by emotion and year. Right: the number of unique nouns per emotion and year, reported per season. The first two collection rounds (2017–S4 and 2018–S1) were excluded because their question format differed from later rounds. Labels such as 2020-S3 indicate the third season of 2020.

is defined as a short narrative describing an individual's personal experience, similar to the life stories in the CoSoWELL corpus (Kyröläinen et al., 2023) and the event descriptions collected in the ISEAR dataset (Troiano et al., 2019). The question was phrased as follows: *What is the most memorable event that has happened to you recently that comes to mind when you think of the following emotion word?* There were no restrictions on the length of the response, and participants were not required to narrate a full story explaining why they felt that way.

In the first two collection rounds (Season 4 in 2017 and Season 1 in 2018), the word "recent" was omitted, and the question was instead: *What is the most memorable event that has happened to you that comes to mind when you think of the following emotion word?* The wording was subsequently refined by adding "recent" to encourage participants to focus on contemporary events. Because this modification may affect the characteristics of the collected data, analyses that compare results across time periods should account for this difference, for example by excluding the first two rounds to ensure comparable conditions.

The emotion words were the seven emotions:

anger, *anxiety*, *disgust*, *trust*, *joy*, *sadness*, and *surprise*. The set was determined with reference to Plutchik's wheel of emotions (Plutchik, 1980), a widely used theoretical framework that organizes basic human emotions into a small number of fundamental categories while capturing their oppositional relationships. We mainly selected emotion terms of moderate intensity from this framework. When an emotion term was not suitable for use in Japanese, we replaced it with a more appropriate alternative. Specifically, "worry" was replaced with *anxiety*, and "acceptance," which is less familiar in everyday Japanese, was replaced with *trust*, which conveys a similar meaning.

Emotion labels in this dataset are self-reported by the participants who experienced the events. Because emotions are inherently subjective experiences, the labels represent the participants' own perceived emotions rather than externally validated annotations. For this reason, inter-annotator agreement was not calculated, as third-party annotators cannot reliably determine the "correct" emotion for a personal episode. Instead, we applied basic quality control procedures to remove invalid responses while preserving the participants' original self-reported labels. This design is consistent with prior datasets that treat emotion labels as self-reported subjective states rather than externally adjudicated annotations, such as the WRIME dataset (Kajiwara et al., 2021), which also collects emotion labels directly from the individuals who experienced the events.

(2) Gender. We asked participants, *Please tell us your gender*. The options were: male, female, and other.

(3) Age. We asked participants, *Please tell us your age group*. The options were: 10s, 20s, 30s, 40s, 50s, 60s, 70s, and other.

3.2. Filtering

A common drawback of crowdsourcing is the potential presence of low-quality responses (Kittur et al., 2008). In our dataset, we also observed clearly noisy entries (e.g., "i'piea'pgka/;lakh/"), which were manually removed. In addition, sensitive information was anonymized to ensure privacy protection and prevent defamation. Personal names, university names, and place names with populations below 20,000 were replaced with the string "***", except when referring to travel destinations. Well-known public figures such as celebrities, politicians, and athletes were not anonymized. Company or product names mentioned in a negative context were also replaced with "***". Furthermore, episodes containing slanderous or discriminatory language, politically or religiously sensitive content, explicit sexual expressions, or other ethically inappropriate descriptions were manually removed during preprocessing. Because the dataset consists of free-form narratives, this filtering necessarily involved qualitative judgment. For example, entries containing explicit insults toward identifiable individuals or groups (e.g., derogatory expressions targeting a particular nationality) or descriptions of explicit sexual relationships between identifiable individuals were excluded. Responses indicating "none" for all seven emotion categories were also excluded.

After applying the above filtering procedures, the remaining dataset sizes by year and emotion are shown in Table 2. Because the data were collected multiple times per year, the number of episodes varies across years. In particular, only one collection wave was conducted in 2017, whereas three waves were conducted in 2024, resulting in smaller totals compared with the other years.

3.3. Participants' Statistics

Figure 2 shows the participants' statistics. The left panel shows gender ratios by year for female, male, and other. The right panel shows the distribution of age groups by year. All values are averaged across the four collection rounds in each year.

Regarding gender, in 2017 when data collection first began, the ratio of males to females was nearly equal. However, over the years, the male ratio has consistently exceeded that of females. Regarding age, most of the participants in the early stages were in their 30s and 40s, but as time passed, the proportion of those in their 50s and 60s increased. These demographic biases constitute one limitation of this dataset, suggesting challenges in creating datasets through crowdsourcing (Behrend et al., 2011). Therefore, when using this dataset, it is necessary to account for these demographic imbalances and, depending on the requirements of the task, apply procedures such as data sampling or stratification.

4. Dataset Analysis

This section profiles textual and lexical characteristics across emotions and periods in our dataset, focusing on text length, noun vocabulary size, and word usage patterns. As noted in Section 3.1, the first two rounds (Season 4 of 2017 and Season 1 of 2018) differ slightly in their elicitation procedure. Therefore, these two rounds are excluded from the analyses in this section.

4.1. Trends in Word Count and Vocabulary Size by Emotion

Figure 3 presents the dataset statistics. The left panel shows the average number of characters per

Year	Anger	Anxiety	Disgust	Joy	Sadness	Surprise	Trust	Total
2017	917	924	908	922	926	920	925	6,442
2018	3,848	3,904	3,813	3,904	3,904	3,899	3,907	27,179
2019	3,676	3,830	3,578	3,832	3,837	3,817	3,831	26,401
2020	3,726	3,881	3,685	3,882	3,875	3,874	3,881	26,804
2021	3,759	3,886	3,691	3,894	3,885	3,890	3,904	26,909
2022	3,764	3,886	3,680	3,893	3,886	3,886	3,894	26,889
2023	3,789	3,874	3,737	3,878	3,881	3,871	3,881	26,911
2024	2,878	2,914	2,864	2,938	2,934	2,916	2,924	20,368
Total	26,357	27,099	25,956	27,143	27,128	27,073	27,147	187,903

Table 2: Number of episodes remaining after filtering, by year and emotion.

episode by emotion and year, reflecting differences in how extensively emotional experiences are described across emotions. The right panel shows the number of unique nouns by emotion and year, reflecting the diversity of events, people, and objects associated with each emotion. All values are aggregated within year. Tokenization for Japanese texts followed MeCab (Kudo, 2005) using the NEologd (Sato et al., 2017) dictionary.

Regarding character count, *joy* is consistently expressed in fewer characters compared to other emotions. On the other hand, *sadness*, *anger*, and *trust* tend to be relatively longer episodes. In terms of noun vocabulary size, *surprise* consistently exhibits a rich variety of lexical items. In contrast, *joy*, *anxiety*, and *trust* tend to show less lexical variation.

Episodes of *sadness* and *anger* in our dataset tend to contain longer episodes, whereas those of *joy* are notably shorter (see the left panel of Figure 3). This pattern aligns with developmental evidence on the negativity bias (Vaish et al., 2008), which suggests that negative emotions receive greater cognitive attention and elaboration than positive ones. Because experiences involving sadness or anger are processed and remembered in more detail, people may provide richer and more extensive descriptions of such episodes, while joyful events are often conveyed in briefer sentences.

On the other hand, the relatively small number of unique nouns for *anxiety* and *trust* appears to reflect their abstract and enduring nature (see the right panel of Figure 3). Rather than being tied to specific events, these emotions are often described as stable psychological states (e.g., "I feel anxious about my future" or "I value family compassion"). Because they refer to generalized experiences rather than concrete episodes, these emotions are typically expressed through a small number of familiar, recurring phrases, leading to a smaller number of unique nouns across episodes.

4.2. Frequency Trends of the Top-10 Most Frequent Words

Figure 4 presents the trends of the 10 most frequent nouns for each emotion. Overall, the plots reveal how lexical salience is shaped not only by the emotional framing but also by external events that vary across time. Below we summarize characteristic tendencies observed for each emotion.

Anger. Words such as "work", "car", and "boss" appear consistently with high frequency. In contrast, terms like "COVID-19" and related words such as "government" and "response" sharply increased during the pandemic period (Season 1 of 2020 to Season 3 of 2021), highlighted in yellow in the *anger* subplot of Figure 4. For instance, one entry in the dataset reads, "The government's response to COVID-19 has been half-hearted", expressing clear frustration toward perceived inaction by authorities. Interestingly, although the number of COVID-19 cases in Japan peaked in 2022, COVID-19-related words were mentioned more frequently in 2021 than in 2022. This discrepancy is likely due to the stricter mobility restrictions repeatedly imposed by the Japanese government in 2021, while very few such restrictions were issued in 2022.

Anxiety. Similar to *anger*, the frequency of "COVID-19" and the related term "infection" increased during the 2020-2021 period, highlighted in yellow in the *anxiety* subplot of Figure 4. For instance, one entry in the dataset states, "The spread of COVID-19 infections", reflecting heightened awareness and anxiety about the expanding outbreak. In contrast, words such as "work" and "future prospect" consistently appeared with high frequency, although the frequency of "future prospect" was relatively lower during the period when words related to COVID-19 were more prevalent. This pattern likely reflects the stronger focus on immediate concerns rather than long-term outlooks during the COVID-19 pandemic.

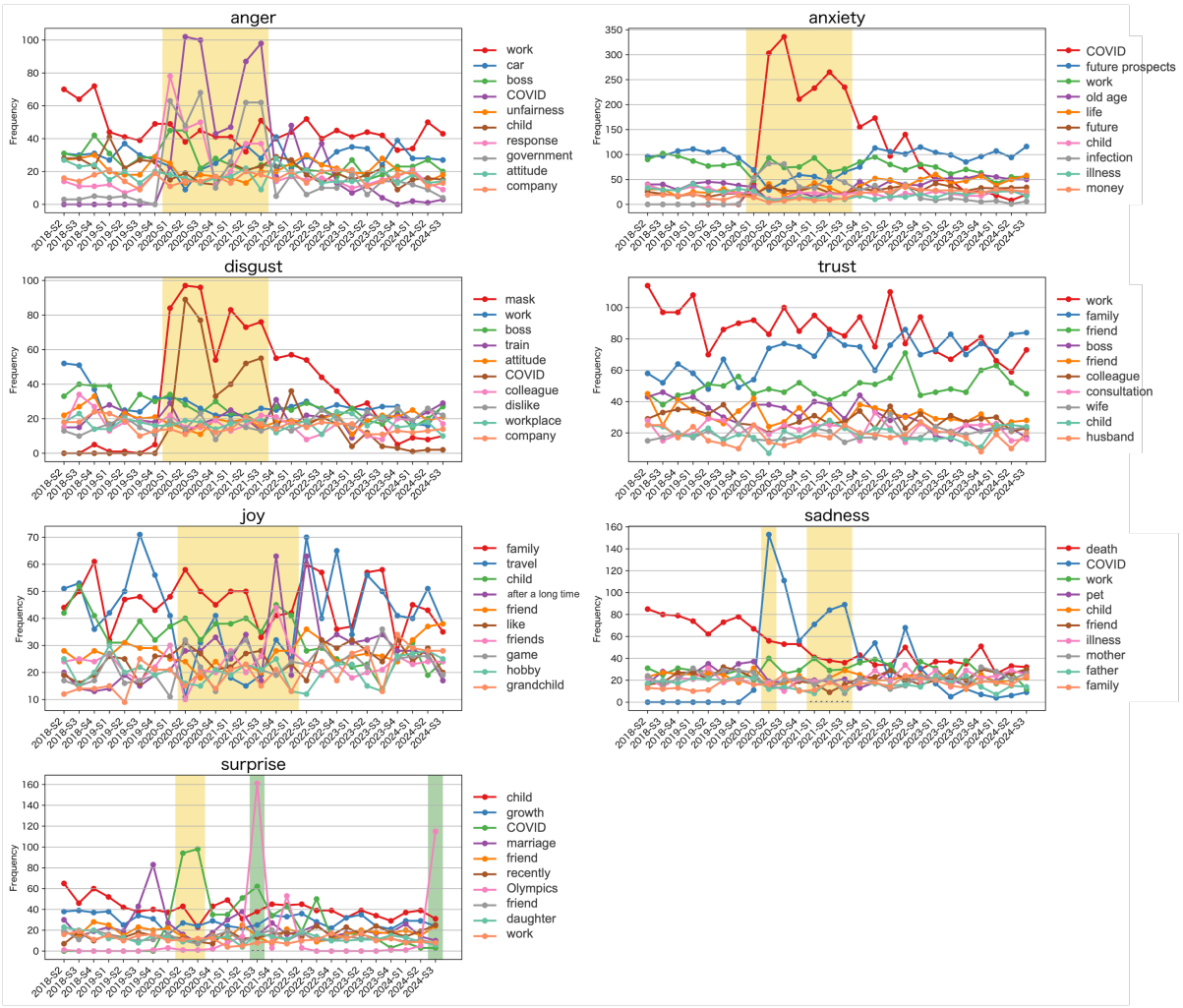

Figure 4: Frequency trends of the top-10 nouns for each emotion. The plots illustrate how salient lexical items rise or fall in usage in correspondence with real-world events (e.g., "the COVID-19 pandemic or the Olympics"). Shaded areas mark the periods that were focused on the analysis in the Subsection 4.2 for each emotion. Note that some nouns are split into multiple words because the original Japanese expressions were translated into English. Labels such as 2020-S3 indicate the third season of 2020.

Disgust. Similar to *anger* and *anxiety*, the frequencies of "COVID-19" and the related term "mask" were high during the 2020–2021 period, highlighted in yellow in the *disgust* subplot of Figure 4. The high frequency of "mask" is particularly characteristic of the *disgust* category, as exemplified by episodes such as "I feel disgust when someone walking nearby isn't wearing a mask." Another distinctive feature is that work-related words, such as "work" and "boss", frequently appear among the top-ranked terms.

Trust. Unlike other emotion categories, the *trust* category shows relatively little variation in word frequency rankings across different periods. Most of the high-frequency words refer to people around the participant who are trusted, with "work" being a notable exception that also appears with high frequency. For example, one entry in the dataset reads, "When I was in trouble, my friend helped me out," expressing interpersonal reliance and gratitude. In contrast to other emotions, words characteristic of the COVID-19 pandemic that consistently rank among the top in other categories do not appear among the top-ranked words in *trust*.

Joy. In contrast to *trust*, the *joy* category exhibits the opposite tendency, with most of the high-frequency words showing substantial variation across periods. For example, the frequency of "travel" declined between Season 2 of 2020 and Season 1 of 2022, highlighted in yellow in the *joy* subplot of Figure 4, corresponding to the period when the Japanese government requested mobility restrictions. By contrast, the expression "after a long time" increased markedly after these restrictions were lifted, sensitively reflecting the social changes of the time. For instance, one entry in the

dataset reads, “We had a great time eating out as a family after a long time,” illustrating a renewed sense of happiness and social reconnection.

Sadness. The timing of increased frequency of the word “COVID-19” closely corresponds to the periods when the Japanese government issued the state of emergency. Specifically, the periods of government restrictions were April 7–May 25, 2020; January 8–March 21, 2021; April 25–June 20, 2021; and July 12–September 30, 2021. The increases in the frequency of “COVID-19” closely coincide with these restriction periods, as shown in the yellow-highlighted area of the *sadness* subplot in Figure 4. For instance, one entry in the dataset states, “Because of COVID-19, I can’t go on trips,” reflecting sadness associated with restricted mobility and loss of ordinary pleasures. This correspondence illustrates how episodic narratives can reveal the emotional impact of major social events in everyday life.

Surprise. The *surprise* category is highly sensitive to trending events. For instance, “COVID-19” frequently appeared in the early stages of the pandemic (Season 2 and 3 of 2020), while “Olympics” spiked at the time of the Olympic Games (Season 3 of 2021 and Season 3 of 2024), highlighted in green in the *surprise* subplot of Figure 4. For example, one entry in the dataset reads, “Japan is winning so many medals at the Olympics,” which expresses astonishment and excitement at the country’s unexpected success. Similarly, “marriage” emerged in Season 4 of 2019, coinciding with a surge of celebrity weddings on November 11, a date considered auspicious for marriage in Japan. It is noteworthy that *surprise* is often associated with very short, event-specific periods in the past.

5. Preliminary Experiments

5.1. Settings

To evaluate whether exposure to our dataset enhances a LLM’s ability to interpret emotions in a time-sensitive manner, we conducted a comparative experiment using three prompting settings implemented with the Gemini 2.5 flash model via the Google Gemini API[3], with the temperature set to 0.3 and top-p to 0.9. These parameters were chosen to balance stability and variation in the model outputs. A non-zero temperature was adopted to avoid fully deterministic decoding and to allow the model to reflect the diversity of human subjective interpretations. Following the silicon sampling approach (Argyle et al., 2023), stochastic decoding was used to better approximate the distribution of human responses in social and behavioral tasks. At the same time, the temperature was kept relatively low to reduce excessive randomness and ensure stable predictions in this classification setting, while a high top-p value was used to avoid overly restrictive token selection. These values were adopted as practical settings for the task rather than parameters optimized through an extensive hyperparameter search.

For each period from Season 1 of 2020 to Season 4 of 2023, the episodes were stratified by emotion label and split into training and test subsets. A fixed set of 100 test instances was drawn per period using stratified sampling, while the remaining episodes served as a pool of candidate training examples for the in-context demonstration conditions. In the **5-shot** and **100-shot** settings, training sentences were sampled proportionally to label frequency and inserted into the prompt as reference pairs of text and emotion label. Specifically, five examples were provided in the **5-shot** setting and 100 examples in the **100-shot** setting. In the **zero-shot** condition, the same test sentences were presented without dataset examples, accompanied only by the corresponding temporal information (year and season). All periods were processed sequentially, and results for each year–season pair were saved individually, along with an aggregated summary across periods. The following prompt template was used in the experiments.

Prompt template used in the experiments

```
You are a careful Japanese text classifier.
↪  Predict one of these categories only:
[ sadness, anxiety, anger, disgust, trust,
↪  surprise, joy ].

Respond ONLY with a JSON object using the schema:
{
  "analysis_summary": "string",
  "distribution": {
    "sadness":  float,
    "anxiety":  float,
    "anger":    float,
    "disgust":  float,
    "trust":    float,
    "surprise": float,
    "joy":      float
  },
  "top_emotion": "string"
}

All values in distribution must be within [0,1]
↪  and sum to 1 (+/-0.01 allowed).
Do not include any text outside the JSON.

Guidance (translated from Japanese):
  - Prioritize meaning over lexical overlap.
  - Use the few shot examples as semantic
    references when provided.
  - Avoid degenerate or uniform distributions.
    Produce plausible probabilities.
  - Keep "analysis_summary" concise and
    informative in Japanese.

[examples as JSONL]
{"text": "Sample text 1", "emotion": "joy"}
{"text": "Sample text 2", "emotion": "sadness"}
... (n shot examples are inserted here)
```

[3] https://ai.google.dev/gemini-api/docs/models, accessed October 14, 2025

```
[input]
[Target Japanese sentence to be classified]
```

The model was instructed to output a normalized probability distribution over the seven emotion categories. We compared the top-predicted emotions among the three prompting settings to assess whether incorporating period-specific episodes from the dataset improved the temporal sensitivity of emotion interpretation. The experimental results are reported based on two evaluation metrics, namely the macro F1 score (F1) and accuracy (Acc).

5.2. Results

Across all periods, the 100-shot condition, which incorporated our dataset as contextual examples, consistently outperformed the zero-shot condition in both macro-F1 and accuracy (Table 3). Overall, macro-F1 and accuracy reached 0.68 and 0.69 in the 100-shot setting, compared to 0.61 and 0.63 in the zero-shot setting, with the 100-shot condition outperforming the zero-shot condition consistently across all periods. The gap was most pronounced in 2020, during the onset of the COVID-19 pandemic, when uncertainty, fear, and social tension likely shaped language use. This indicates that the model benefited from our dataset, which helped it capture time-sensitive emotional nuances.

On the other hand, the 5-shot setting showed little difference from the zero-shot setting in our experiments, suggesting that a small number of in-context examples was insufficient to substantially improve performance for this task. A previous study has reported that increasing the number of in-context examples can further enhance model performance (Abbas et al., 2024), and that many-shot prompting may function similarly to pseudo fine-tuning under certain conditions. Our results are consistent with this line of research in that limited few-shot examples alone did not lead to clear improvements.

For *disgust*, cases correctly classified only by the 100-shot model were more common than for any other emotion, while cases correctly classified only by the zero-shot model were rare (Table 4). Most episodes labeled as *disgust* were misclassified as *anger* or *anxiety* under the zero-shot condition. Many of these cases involved texts related to mask-wearing during the COVID-19 pandemic. For example, the Season 4 of 2020 entry “People not wearing masks on the train” was predicted as *anxiety* instead of *disgust* as shown in Table 5 (1). This suggests that without contextual grounding from period-specific examples, the model tended to generalize public-norm violations as expressions of fear or frustration rather than moral aversion.

Period	100-shot		5-shot		Zero-shot	
	F1	Acc	F1	Acc	F1	Acc
2020-S1	0.81	0.82	0.68	0.69	0.67	0.66
2020-S2	0.70	0.70	0.57	0.59	0.57	0.57
2020-S3	0.70	0.70	0.64	0.65	0.64	0.65
2020-S4	0.65	0.66	0.57	0.61	0.56	0.57
2021-S1	0.71	0.72	0.65	0.67	0.64	0.66
2021-S2	0.63	0.64	0.58	0.60	0.53	0.57
2021-S3	0.66	0.67	0.69	0.70	0.63	0.63
2021-S4	0.69	0.69	0.56	0.59	0.53	0.57
2022-S1	0.70	0.70	0.64	0.65	0.65	0.66
2022-S2	0.60	0.63	0.65	0.67	0.58	0.61
2022-S3	0.67	0.69	0.62	0.64	0.66	0.68
2022-S4	0.76	0.77	0.72	0.72	0.73	0.73
2023-S1	0.64	0.65	0.62	0.63	0.61	0.63
2023-S2	0.69	0.69	0.60	0.63	0.61	0.64
2023-S3	0.65	0.65	0.58	0.59	0.57	0.59
2023-S4	0.59	0.61	0.59	0.61	0.55	0.57
Overall	**0.68**	**0.69**	**0.62**	**0.64**	**0.61**	**0.63**

Table 3: Comparison of performance metrics across all periods under zero-shot, 5-shot, and 100-shot settings. Labels such as 2020-S3 indicate the third season of 2020. Acc means accuracy and F1 means macro F1 score.

Emotion	Both correct	100s only	Zs only	Both error
anger	0.51	0.05	0.10	0.33
anxiety	0.86	0.03	0.07	0.04
disgust	0.34	0.22	0.01	0.43
trust	0.55	0.16	0.06	0.24
joy	0.82	0.05	0.04	0.09
sadness	0.67	0.07	0.04	0.23
surprise	0.31	0.17	0.01	0.51

Table 4: Prediction outcome proportions by gold emotion. Values are rowwise proportions. 100s denotes 100-shot and Zs denotes zero-shot condition.

In contrast, under the 100-shot condition, a greater number of errors were observed for *anxiety*, compared to the zero-shot condition. For example, one such case is the text “I suddenly gained weight,” (see (2) of Table 5) where the reasoning provided by the LLM included statements such as “a sense of anxiety can be inferred” in both the 100-shot and zero-shot conditions. In many other examples as well, the model’s explanations indicated the simultaneous presence of both *anxiety* and *sadness*, suggesting that the LLM does not draw a clear boundary between these two emotions.

Finally, *surprise* and *disgust* showed a number of common errors between the two conditions. For example, the text “The increase in COVID-19 pos-

index	Text	Gold label	Predicted	
			100-shot	Zero-shot
(1)	People not wearing masks on the train	*disgust*	*disgust*	*anxiety*
(2)	I suddenly gained weight	*sadness*	*anxiety*	*anxiety*
(3)	The increase in COVID-19 positive cases	*surprise*	*anxiety*	*anxiety*
(4)	A neighboring country's gesture of reconciliation	*disgust*	*trust*	*trust*

Table 5: Examples cited in the error analysis (originally in Japanese, translated into English).

itive cases" was classified as *anxiety* under both conditions, although the gold label was *surprise* as shown in Table 5 of (3). When an episode conveys a clearly positive or negative emotional tone, such as fear or concern, even if it was originally labeled as *surprise*, the model tends to prioritize conventional polarity cues over the intended category, leading to misclassification. Similarly, the text "A neighboring country's gesture of reconciliation" was labeled as *trust* in both conditions, while the gold label was *disgust* as shown in Table 5 (4). Accurately interpreting such emotions would require contextual knowledge about the writer's background or stance, highlighting the inherent difficulty of emotion classification from single-sentence inputs. A practical way to address this issue would be to incorporate approaches that estimate emotion distributions or model uncertainty, as recently suggested (Wu et al., 2024).

Taken together, these results demonstrate that incorporating period-specific examples from the dataset enhances the temporal sensitivity and contextual precision of emotion interpretation in Japanese texts. While zero-shot prompting captures general emotional categories, it often fails to recognize culture- and time-specific nuances, particularly for emotions such as *disgust* and *surprise*, which rely heavily on social context. These findings highlight the usefulness of our dataset in providing temporally grounded emotional episodes that help language models better capture socially contextualized expressions of emotion.

6. Conclusion and Future Work

We introduced NAIST LIFE STORY, a Japanese dataset that enables identifying a time-aware relationship between emotions and texts. This dataset consists of episodes elicited by seven emotions (anger, anxiety, disgust, trust, joy, sadness, and surprise), collected quarterly, and includes the age and gender of the participants who provided them. The descriptive analyses and the preliminary experiment indicate that the dataset can surface changes in associations between emotions and texts that are consistent with social phenomena. We plan to continue the data collection for more than a decade to build a long-term record of emotional language use. Future work includes continued long-term collection, expansion to additional languages to support cross cultural comparison, and deeper modeling of temporal dynamics with stronger evaluation protocols.

Limitations

Language and cultural scope The dataset is limited to Japanese. Because emotional expression depends on language and culture, the patterns observed in this dataset may not directly generalize to other languages. In particular, the interpretation of some emotions may depend on culturally shared background knowledge. Future work should apply the same data collection framework to other languages to examine cross-cultural similarities and differences.

Temporal coverage The dataset covers seven years, which is shorter than the time spans examined in many studies of semantic change that analyze century-scale corpora (Hamilton et al., 2016; Xu et al., 2021). Therefore, the dataset is not designed to capture very long-term historical shifts. Instead, the repeated data collection at short intervals allows the analysis of relatively fine-grained changes in what people associate with specific emotions in everyday life. Future work should clarify how such short-term changes relate to longer-term semantic change.

Ambiguity between emotion categories The error analysis indicates that some emotions remain difficult to distinguish. For example, *disgust* was sometimes classified as *anxiety* or *sadness*, and *sadness* was sometimes confused with *anxiety*. These results suggest that neighboring emotions may share lexical or contextual features.

One possible reason for this ambiguity is the structure of the dataset itself. In the present dataset, participants describe concrete episodes associated with an emotion. In contrast, many NLP resources assign emotion labels directly to short texts or sentences. For example, the GoEmotions dataset annotates Reddit comments with fine-grained emo-

tion categories (Demszky et al., 2020). Similarly, emotion lexicons such as the NRC Emotion Lexicon (Mohammad and Turney, 2013) or WordNet-Affect (Strapparava and Valitutti, 2004) associate individual words with emotion categories. Because these resources rely on different units of analysis (episodes, sentences, or words), the similarity structure between emotions may differ across datasets. For instance, emotions that are clearly separated in lexicon-based resources may appear closer when people recall real-life episodes, where multiple emotional states can co-occur or overlap within the same narrative.

A promising direction for future work is therefore to compare the structure of emotion similarity across different types of resources. Such comparisons could examine whether the distances between emotions observed in episode-based narratives are consistent with those derived from sentence-level emotion datasets or lexicon-based emotion resources. This type of analysis could help clarify whether the observed confusion reflects properties of the dataset, characteristics of narrative emotion expression, or broader patterns in computational representations of emotion.

Scope of the LLM experiment The LLM experiment in this study is limited in several respects. It evaluates only a single model (Gemini 2.5 Flash), uses a relatively small evaluation set (100 instances per period). Future work should test multiple models and use larger evaluation sets to better assess the practical usefulness of the dataset.

Prompt design and temporal context The current prompting conditions include only basic temporal information such as year and season. They do not incorporate more explicit contextual descriptions of major social events during the target period. Future work could explore richer prompt designs that include structured temporal information or background knowledge about important events.

Limited gains with small-scale prompting The improvement observed in the 100-shot setting was modest. Therefore, the present results should be interpreted as preliminary evidence rather than definitive proof of the effectiveness of using the NAIST LIFE STORY dataset in prompting.

Recent studies suggest that increasing the number of in-context examples can substantially improve model performance, sometimes approaching the effect of lightweight fine-tuning (Abbas et al., 2024). From this perspective, the prompting configurations explored in this study may still be relatively small.

A natural next step is therefore to investigate how performance changes when substantially larger numbers of examples are included in the prompt. In particular, it would be informative to examine how performance scales with the number of in-context examples and to identify the point at which additional examples no longer yield meaningful improvements. Such an analysis would help estimate the amount of data required for effective prompting in this task.

Ethical Considerations

All participants provided informed consent prior to data collection. The dataset contains no personally identifiable information. Data were collected through a commercial crowdsourcing platform, and workers were compensated at a fair market rate. The study followed the ethical guidelines for research involving human subjects.

Acknowledgements

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on “Integrated Health Care System” Grant Number JPJ012425.

Bibliographical References

Zaheer Abbas, Rishabh Agarwal, Ankesh Anand, Feryal Behbahani, Bernd Bohnet, Stephanie Chan, Eric Chu, John Co-Reyes, Aleksandra Faust, Hugo Larochelle, Azade Nova, Luis Rosias, Avi Singh, Biao Zhang, and Lei Zhang. 2024. Many-shot in-context learning. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pages 76930–76966. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Tara S. Behrend, David J. Sharek, Adam W. Meade, and Eric N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43(3):800–813.

British Library. n.d. National life stories. https://www.bl.uk/about/projects/national-life-stories. Accessed October 14, 2025.

Charles P. Davis. 2023. Emergence of covid-19 as a novel concept shifts existing semantic spaces. *Cognitive Science*, 47(1):e13237.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. 2018. Diachronic degradation of language models: Insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Melbourne, Australia. Association for Computational Linguistics.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. Wrime: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456, New York, NY, USA. Association for Computing Machinery.

Anna Koufakou and Elijah Nieves. 2025. Review of recent emotion-annotated text corpora and resources. *Language Resources and Evaluation*.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. https://taku910.github.io/mecab/. Accessed: 2025-10-14.

Aki-Juhani Kyröläinen, James Gillett, Megan Karabin, Ranil Sonnadara, and Victor Kuperman. 2023. Cognitive and social well being in older adulthood: The cosowell corpus of written life stories. *Behavior Research Methods*, 55:2885–2909.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, and Evangelia Adamou. 2014. Documenting and researching endangered languages: The pangloss collection. *Language Documentation & Conservation*, 8:119–135.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval 2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Deepa Narayan, Raj Patel, Kai Schafft, Anne Rademacher, and Sarah Koch-Schulte. 2000. *Voices of the poor. Can anyone hear us?* Oxford University Press.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum Associates.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. Analyzing semantic change through lexical replacements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Robert Plutchik. 1980. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Masato Sato, Takaaki Sato, and Manabu Okumura. 2017. Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese). In *Proceedings of the 23rd Annual Meeting of the Association for Natural Language Processing*, pages NLP2017–B6–1, Tokyo, Japan. The Association for Natural Language Processing.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: An affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1083–1086.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Enrica Troiano, Laura Oberländer, Maximilian Wegge, and Roman Klinger. 2022. x-envent: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1365–1375, Marseille, France. European Language Resources Association.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event focused emotion corpora for german and english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Amrisha Vaish, Tobias Grossmann, and Amanda L. Woodward. 2008. Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological bulletin*, 134 3:383–403.

Wen Wu, Bo Li, Chao Zhang, Chung-Cheng Chiu, Qiujia Li, Junwen Bai, Tara Sainath, and Phil Woodland. 2024. Handling ambiguity in emotion: From out-of-domain detection to distribution estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2078–2093, Bangkok, Thailand. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Aotao Xu, Jennifer E. Stellar, and Yang Xu. 2021. Evolution of emotion semantics. *Cognition*, 217:104875.

7. Language Resource References

Kajiwara, Tomoyuki and Chu, Chenhui and Takemura, Noriko and Nakashima, Yuta and Nagahara, Hajime. 2021. *WRIME: Japanese Social Media Posts with Writer and Reader Emotion Intensities*. Eight emotions, writer and reader perspectives.

Michailovsky, Boyd and Mazaudon, Martine and Michaud, Alexis and Guillaume, Severine and François, Alexandre and Adamou, Evangelia. 2014. *The Pangloss Collection*. Language Documentation & Conservation. Oral texts with transcripts and translations for endangered and other languages.

Mohammad, Saif M. and Turney, Peter D. 2013. *NRC Emotion Lexicon*. NRC Canada. English word-emotion associations created by crowdsourcing.

Pennebaker, James W. and Booth, Roger J. and Boyd, Ryan L. and Francis, Martha E. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates. PID https://www.liwc.app.

Troiano, Enrica and Padó, Sebastian and Klinger, Roman. 2019. *DeISEAR: Event-focused Emotion Corpora*. Short event descriptions with emotion labels.