From Lived Experience to Text: A Corpus in Peruvian Spanish of Household and Civic Narratives

Anonymous ACL submission

Abstract

We present a sentence-level corpus derived from lived experiences and civic events in South America, designed for qualitative and quantitative studies of physical common reasoning. The corpus focuses on everyday, observable knowledge about objects, materials, bodies, environments, and their causal interactions. Sentence prompts were elicited from natural interactions, as well as from public-interest topics affecting speakers of Peruvian Spanish. To support reasoning studies, the dataset also includes physically implausible variants, typically created by making minor changes to otherwise correct sentences.

1 Objective

Our objective is to construct a sentence-level corpus reflecting (i) personal and family experiences and (ii) locally relevant civic events, with a focus on physical common sense expressed through observable phenomena and concrete actions.

2 Source Population and Scope

Authors adapted sentences from naturally occurring speech and reconciled differences by consensus. These included personal experiences and experiences of family and friends. The sentences are drawn from public-interest topics in Lima, Peru (e.g., local traditions, bank crimes, or the conduct of public officials). The was collected for a month, between August 1 and September 1, 2025.

3 Inclusion and Exclusion Criteria

Inclusion. Sentences were included if they described concrete experiences or observations, specified procedures or actions such as hygiene or food preparation, or encoded interactions between objects and materials, including safety constraints.

Exclusion. Sentences were excluded if they expressed opinions or value judgments without observable evidence, contained private identifiers, or were speculative, unverifiable, or unrelated to physical common sense.

4 Data Processing

De-identification. Names, addresses, and direct identifiers were removed and replaced with role-based or categorical placeholders.

Normalization. To preserve diversity, tense and punctuation were not standardized but instead left reflective of colloquial speech.

5 Intended Use and Availability

The corpus supports both qualitative coding of everyday physical reasoning and quantitative evaluations (e.g., label distributions, agreement rates). It is intended for benchmarking models on grounded common-sense assertions. Redistribution should preserve de-identification and adhere to this usage policy.