Building a Humor-Type Corpus from Caption Contest Data

Leila Habibi
LING7800 Corpus Analysis
Instructor: Kyle Gorman
Spring 2025

Project Report

Building a Humor-Type Corpus from Caption Contest Data

LING7800 Corpus Analysis – Leila Habibi

1. Project Overview

This project involved the manual annotation of captions from the New Yorker Caption Contest dataset compiled by NextML. The dataset comprises crowd-sourced captions submitted to approximately 390 caption contests, each associated with a single-panel cartoon. For this project, 200 contests were randomly selected, and the five top-ranked captions per contest were extracted, resulting in a corpus of approximately 1,000 caption–cartoon pairs.

The goal was to create a humor-type corpus by categorizing each caption into one of five predefined humor categories:
- Pun / Wordplay
- Irony / Sarcasm
- Absurdity / Nonsense
- Cultural Reference
- Expectation Violation

Each caption was annotated along with its contest ID, with an option for annotators to add notes where categorizations were ambiguous. Humor categories were numerically coded (001 to 005) to enable future computational analysis.

2. Annotation Process

Two annotators independently labeled all captions according to the humor-type guidelines. The annotation schema was based on the following category definitions:

Code	Humor Type	Definition
001	Pun / Wordplay	Clever manipulation of language or sounds
002	Irony / Sarcasm	Mockery or saying the opposite of literal meaning
003	Absurdity / Nonsense	Illogical, surreal humor
004	Cultural Reference	Based on external knowledge or cultural cues
005	Expectation Violation	Surprise twist that subverts the familiar

The annotation guidelines were developed to ensure shared understanding, and several example-based discussions were held to resolve initial ambiguities.

3. Inter-Annotator Agreement (IAA)

To evaluate annotation reliability, Cohen's Kappa (κ) was computed based on the contingency table of the two annotators' labels:

		Anu				
	CODES	**001**	**002**	**003**	**004**	**005**
	001	30	3	-	1	-
	002	2	16	1	-	4
Leila	**003**	-	6	5	-	3
	004	2	3	1	6	-
	005	3	7	3	1	3

Agreement Results:
- Observed Agreement (Po): 0.60
- Expected Agreement (Pe): 0.248
- Cohen's Kappa (κ): 0.468

This κ score reflects moderate agreement on the Landis and Koch qualitative scale.

4. Adjudication Phase

Following the IAA analysis, an adjudication phase was conducted where annotators discussed disagreements and resolved category assignments to create a gold-standard corpus. This adjudicated dataset will be used for developing a tagger as the course project for Computational Methods II, on which I expand more in section 7.

5. Challenges Encountered

- Category Overlap: Some captions could validly fit into multiple humor categories, making annotation decisions difficult.
- Ambiguity: Certain captions were inherently vague or relied on cultural knowledge that may not be universally shared, leading to disagreements.

- Annotation Fatigue: The high cognitive load required to consistently differentiate humor types may have contributed to minor inconsistencies over time.

6. Scientific Impact and Future Directions

The annotated corpus produced in this project offers a valuable resource for computational humor detection, multimodal meaning analysis, and pragmatic studies of humor. Since the original dataset includes crowd-sourced funniness ratings, future work can investigate potential correlations between humor types and audience perceptions.

Future Improvements:
- Increasing the number of annotators to improve reliability.
- Providing additional training with edge-case examples.
- Refining humor type definitions to reduce overlap and/or ambiguity.

7. Integration with LING83800 Project

In this project, a manual annotation technique was applied to categorize approximately 1,000 captions into predefined humor types. As a complementary effort in my Methods in Computational Linguistics II course, I extended this work by training and evaluating an automatic tagger capable of performing the humor classification task. The goal of the Methods II project was to explore the potential of computational models to replicate the manual annotation process and to assess their performance on this nuanced linguistic task. While the corpus annotation and the tagger training are closely related, they serve distinct purposes: the manual annotation project focused on building a high-quality humor-type corpus, whereas the Methods II project aimed to develop and evaluate an automated system using that corpus as data.

8. Appendices

The annotation sheets containing the cartoons, their corresponding captions, and the assigned humor type codes are provided as appendices to this report.

9. References

NextML. (n.d.). New Yorker Caption Contest Data. Retrieved from https://nextml.github.io/caption-contest-data/

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555–596.

LING78000: Corpus Analysis. (n.d.). Inter-annotator Agreement [PDF]. Retrieved from course materials.