Experimental versus In-Corpus Variation in Referring Expression Choice

Fafa Same

(Joint work with T. Mark Ellison)

Utrecht NLP Group

December 14, 2023

Referring is (most of the times) a non-deterministic task

Text one

Homer Jay Simpson$_{properN}$ (born May 12 1956) is the main protagonist and one of the five main characters of The Simpsons series (or show). **He**$_{pronoun}$ is the spouse of Marge Simpson and father of Bart, Lisa and Maggie Simpson. **Homer**$_{properN}$ is overweight (said to be 240 pounds), lazy, and often ignorant to the world around **him**$_{pronoun}$.

Text two

Homer Jay Simpson$_{properN}$ (born May 12 1956) is the main protagonist and one of the five main characters of The Simpsons series (or show). **Homer**$_{properN}$ is the spouse of Marge Simpson and father of Bart, Lisa and Maggie Simpson. **He**$_{pronoun}$ is overweight (said to be 240 pounds), lazy, and often ignorant to the world around **him**$_{pronoun}$.

However ...

Models for predicting referring expression forms (RFs) are generally evaluated against corpora of written texts, offering a single correct response in the given context.

However ...

Models for predicting referring expression forms (RFs) are generally evaluated against corpora of written texts, offering a single correct response in the given context.

And we don't like this! ☹

Aim of this study

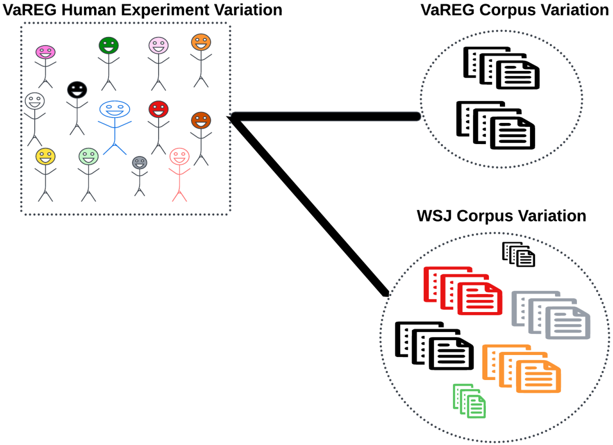
To explore how well the variation seen in a large corpus, like the Wall Street Journal (WSJ), can function as a proxy for variation found in experiments.

Presentation Outline

- Previously on this topic
- Current study
- Where do these inferred distributions come from?
- Hypotheses
- Results
- Conclusion and Discussion

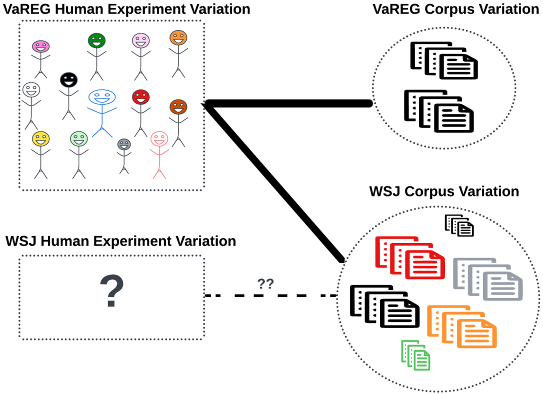
Previously on this topic: my presentation in March 2023

My presentation in March:



- Constructing distributions of RE variation from corpora
- Compare variation in human behavior (Castro Ferreira et al., 2016a) with variation found within a text corpus
- Showing that the in-corpus WSJ distributions matched the human VaREG distributions seen when multiple speakers choose RE forms for the same referent.

My presentation in March:

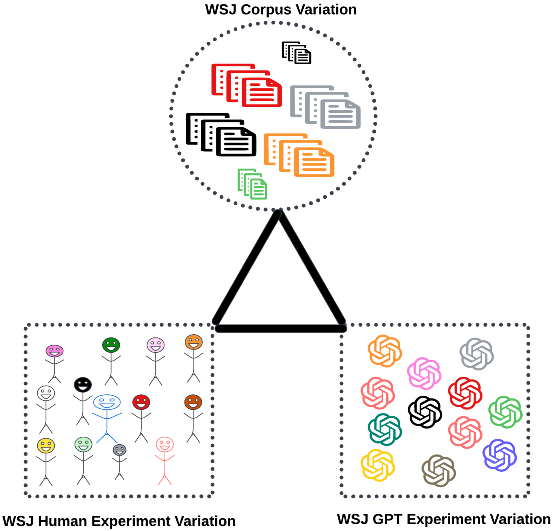

VaREG Human Experiment Variation
VaREG Corpus Variation
WSJ Corpus Variation
WSJ Human Experiment Variation
?
??

- Constructing distributions of RE variation from corpora
- Compare variation in human behavior with variation found within a text corpus
- Showing that the in-corpus WSJ distributions matched the human VaREG distributions seen when multiple speakers choose RFs for the same referent.
- But sadly, we did not compare the inferred within-corpus WSJ distributions with the WSJ human variations

Current study

Current Study

GOAL: How well the corpus itself can model the variation found when multiple informants (either human participants or LLMs) choose referential expressions in the same contexts.



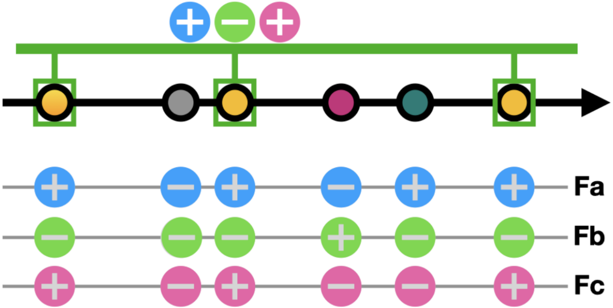
Where do these probability distributions come from?

Within-corpus distributions

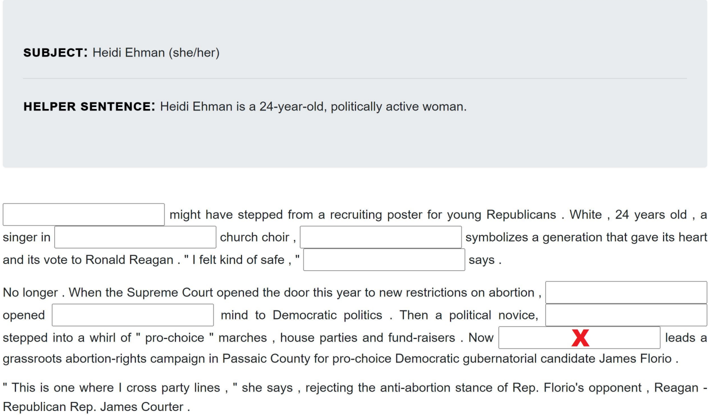
Wall Street Journal (WSJ):
500 documents (30,500 REs)

3 RE categories:
Pronoun, Proper Name, Description

Feature-value categories:
Grammatical role, antecedent form, animacy, sentential recency, paragraph transition



Example



SUBJECT: Heidi Ehman (she/her)

HELPER SENTENCE: Heidi Ehman is a 24-year-old, politically active woman.

□ might have stepped from a recruiting poster for young Republicans . White , 24 years old , a singer in □ church choir , □ symbolizes a generation that gave its heart and its vote to Ronald Reagan . " I felt kind of safe , " □ says .

No longer . When the Supreme Court opened the door this year to new restrictions on abortion , □ opened □ mind to Democratic politics . Then a political novice, □ stepped into a whirl of " pro-choice " marches , house parties and fund-raisers . Now □ X leads a grassroots abortion-rights campaign in Passaic County for pro-choice Democratic gubernatorial candidate James Florio .

" This is one where I cross party lines , " she says , rejecting the anti-abortion stance of Rep. Florio's opponent , Reagan - Republican Rep. James Courter .

Original RE: she

Produced REs: Heidi Ehman, Ehman, the 24-year-old, the former Republican, politically active woman, she, they

Inferred Human/GPT Distributions

Inferred human distributions:

1. Crowd-sourcing experiment
2. A total of 50 documents, and 414 referential gaps
3. 5 lists, each with 10 items

city-country	human	org	other
10	20	10	10

4. Each gap is filled by 20 participants (a total of 100 participants)
5. ≈8300 judgments
6. Fafa: Manual annotation of RE forms

Inferred Human/GPT Distributions

Inferred human distributions:

1. Crowd-sourcing experiment
2. A total of 50 documents, and 414 referential gaps
3. 5 lists, each with 10 items

city-country	human	org	other
10	20	10	10

4. Each gap is filled by 20 participants (a total of 100 participants)
5. ≈8300 judgments
6. Fafa: Manual annotation of RE forms

Inferred GPT distributions:

1. OpenAI's GPT-4 (model=gpt-4)
2. Prompt similar to the instructions given to humans (with minor modifications)
3. Same lists as the human experiment
4. Each list was run 20 times
5. After the experiment, ChatGPT was used to annotate (1) RE form, and (2) whether the generated REs accurately referred to the intended referents.
6. 1-2 hours of manual correction

Inferred Human/GPT Distributions

Inferred human distributions:

1. Crowd-sourcing experiment
2. A total of 50 documents, and 414 referential gaps
3. 5 lists, each with 10 items

city-country	human	org	other
10	20	10	10

4. Each gap is filled by 20 participants (a total of 100 participants)
5. ≈8300 judgments
6. Fafa: Manual annotation of RE forms

Inferred GPT distributions:

1. OpenAI's GPT-4 (model=gpt-4)
2. Prompt similar to the instructions given to humans (with minor modifications)
3. Same lists as the human experiment
4. Each list was run 20 times
5. After the experiment, ChatGPT was used to annotate (1) RE form, and (2) whether the generated REs accurately referred to the intended referents.
6. 1-2 hours of manual correction

Achieved probability distributions over RE forms for
(1) each reference slot, and (2) each feature-category

Hypotheses

Hypotheses

1. Human variation in RE forms is lower for REs belonging to the same feature category.

Hypotheses

1. Human variation in RE forms is lower for REs belonging to the same feature category.
2. In-corpus variation and human experimental variation are more similar when the categories are aligned than when they are not.

Hypotheses

1. Human variation in RE forms is lower for REs belonging to the same feature category.
2. In-corpus variation and human experimental variation are more similar when the categories are aligned than when they are not.
3. The experiment will show greater use of pronouns than the original corpus.

Hypotheses

1. Human variation in RE forms is lower for REs belonging to the same feature category.
2. In-corpus variation and human experimental variation are more similar when the categories are aligned than when they are not.
3. The experiment will show greater use of pronouns than the original corpus.
4. The experiment will show fewer use of descriptions than the original corpus.

Hypotheses

1. Human variation in RE forms is lower for REs belonging to the same feature category.
2. In-corpus variation and human experimental variation are more similar when the categories are aligned than when they are not.
3. The experiment will show greater use of pronouns than the original corpus.
4. The experiment will show fewer use of descriptions than the original corpus.
5. GPT distributions align more closely with corpus distributions than human distributions do.

Hypotheses

1. Human variation in RE forms is lower for REs belonging to the same feature category.
2. In-corpus variation and human experimental variation are more similar when the categories are aligned than when they are not.
3. The experiment will show greater use of pronouns than the original corpus.
4. The experiment will show fewer use of descriptions than the original corpus.
5. GPT distributions align more closely with corpus distributions than human distributions do.
6. The number of descriptions produced by the GPT-4 models exceed those produced by humans.

Results

$\mathcal{H}_1$ Variation patterns are more similar within categories than between them

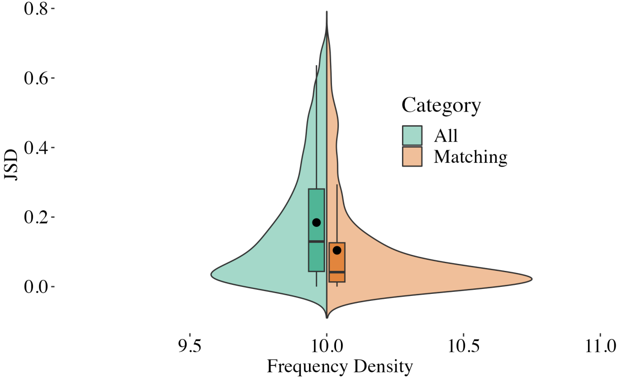


Figure 1: Mean: dots within the box plots. Median: The center line in the boxes (higher in green). Quartiles: Box boundaries, showing the middle 50% of data. Whiskers: approximately 95% of the data range.

$\mathcal{H}_1$ Variation patterns are more similar within categories than between them

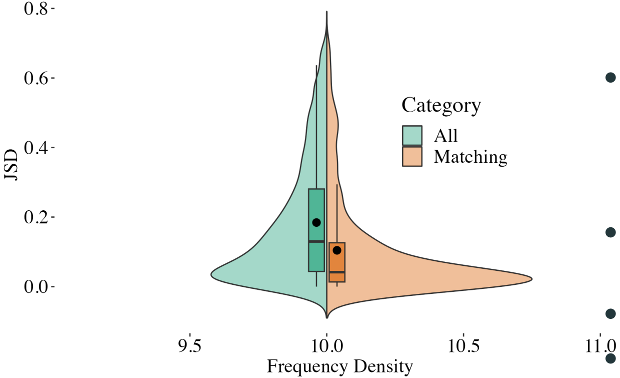
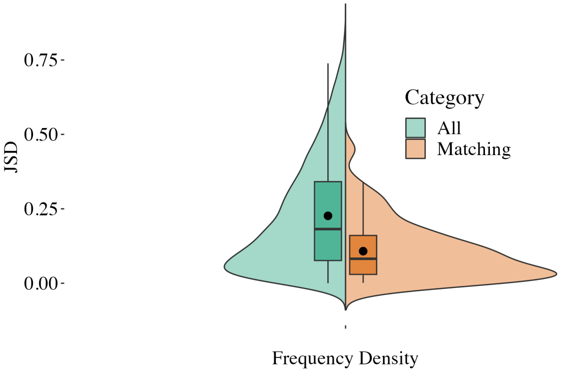


Figure 1: Mean: dots within the box plots. Median: The center line in the boxes (higher in green). Quartiles: Box boundaries, showing the middle 50% of data. Whiskers: approximately 95% of the data range.

- Jensen-Shannon Divergence (JSD) measures similarity between probability distributions. **Lower JSD** indicates **higher similarity** between slots.
- For each pair of unique REs in human data.
- Green (All Pairs): Higher variation in JSD.
- Orange (Same Category): a strong reduction in JSDs.
- Median JSD for all pairs > the 3rd quartile for matching pairs.

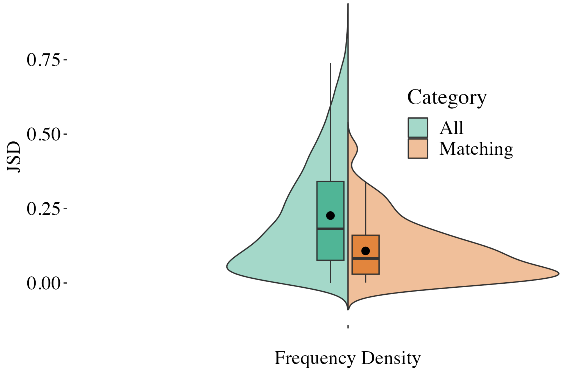
$\mathcal{H}_2$ In-corpus and experimental variation more similar when aligned

Split violin plot and box plot



$\mathcal{H}_2$ In-corpus and experimental variation more similar when aligned

Split violin plot and box plot



- JSDs of corpus and experimental form distributions with identical (Matching) and arbitrary (matching or mismatching) feature-value combinations (All)
- Much larger JSDs for the non-matching categories.
- The median in the randomised comparisons is around twice that of the matched comparisons.

$\mathcal{H}_3$ More pronouns in the experimental results
$\mathcal{H}_4$ Fewer definite descriptions in the experimental results

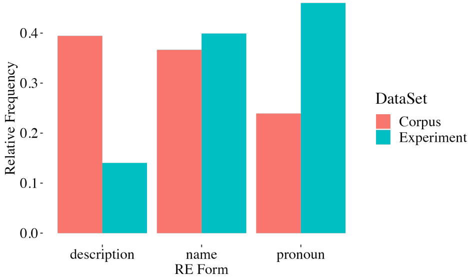


Figure 2: Relative frequencies of different forms in corpus and experimental results.

$\mathcal{H}_3$ More pronouns in the experimental results
$\mathcal{H}_4$ Fewer definite descriptions in the experimental results

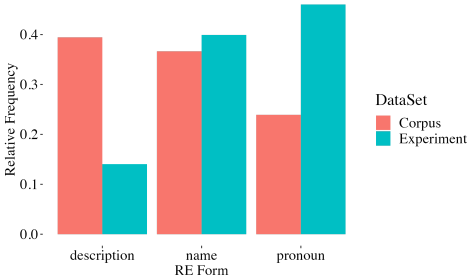


Figure 2: Relative frequencies of different forms in corpus and experimental results.

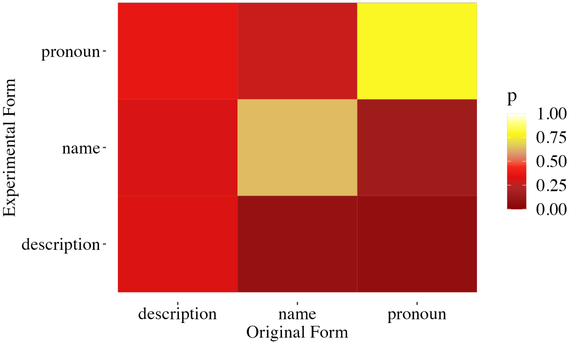


Figure 3: strong correlation between the pronominal realisation in Corpus and in Human. Substantial agreement for proper names between the two (central column). Inconsistent realisation for descriptions.

$\mathcal{H}_5$ GPT distributions more similar to corpus, than to human distributions

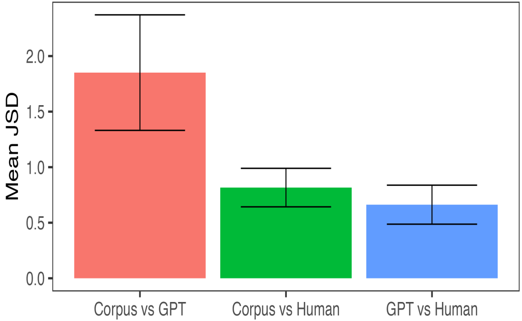


Figure 4: Mean JSDs comparing Corpus distributions, human distributions, and GPT-4 distributions.

$\mathcal{H}_5$ GPT distributions more similar to corpus, than to human distributions

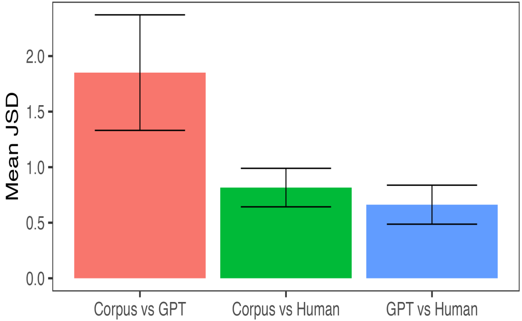


Figure 4: Mean JSDs comparing Corpus distributions, human distributions, and GPT-4 distributions.

As shown by the non-overlapping standard-error error-bars, there a significant difference between the mean JSDs of the Corpus and GPT on the one hand (red) and Corpus and Human on the other (green), with Corpus exhibiting variation more similar to that of human experimental participants. In fact, the Corpus-Human distance is only slightly larger than the GPT-Human distance, showing that the Corpus distributions form almost as good a model of human-experimental variation as does the LLM.

$\mathcal{H}_6$ GPT would produce more descriptive REs than human participants.

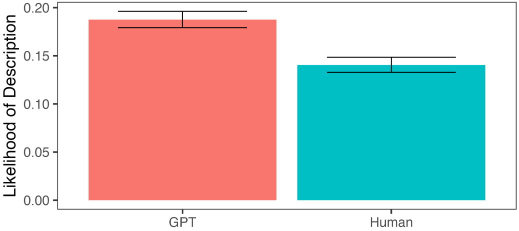


Figure 5: The likelihood of description (vs pronoun or name) use in experimental results: GPT vs Human. The error bars show the Beta-Bernoulli 95confidence intervals, and do not overlap, so the difference in results is unlikely to be due to chance.

Conclusion and Discussion

Conclusion

AIM: To explore how well the variation seen in a large corpus, like the WSJ, can function as a proxy for variation found in experiments.

- Distributions defined by the same feature-value categories are much more similar than those defined by arbitrary matching ($\mathcal{H}_2 1$ & $\mathcal{H}_2$) → evidence of a common cause at play in conditioning these distributions.
- We observe different reference strategies in corpus as opposed to the experimental setup.
- More, but not all, of the variation in the Human are matched in GPT. However substantial aspects of human variation were captured by the Corpus model but not by GPT.

References i

References

Castro Ferreira, T., E. Krahmer, and S. Wubben (June 2016a). "Individual Variation in the Choice of Referential Form". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 423–427.

Castro Ferreira, T., E. Krahmer, and S. Wubben (Aug. 2016b). "Towards more variation in text generation: Developing and evaluating variation models for choice of referential form". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 568–577.

Recasens, M., E. Hovy, and M. A. Martí (May 2010). "A Typology of Near-Identity Relations for Coreference (NIDENT)". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Formulas

Measure of entropy: X probability of RF in a given context, n = number of classes.

$$H(x) = -\sum_{i=1}^{n=3} \frac{p(x_i)\log(p(x_i))}{\log(n)} \tag{1}$$

where x corresponds to references in the current gap, and n=3 is the number of different forms annotated. The probability of having a RF i express token x is given by p(xi).

Jensen Shannon divergence:

The Kullback-Liebler Divergence expresses degrees of difference between distributions. It can be thought of as the average amount of extra information which must be supplied to represent an item x occurring with relative frequency p(x), if it was expected with frequency q(x).

$$KL(p||q) = \sum_{x} p(x)\log_2 \frac{p(x)}{q(x)} \tag{2}$$

Given two distributions p and q, the Jensen-Shannon divergence metric JSD is the average of the KLD measures from a midpoint distribution $r = \frac{p+q}{2}$ to p and to q. This measure has the desirable property of being 0 for identical distributions, and 1 for maximal divergence.