# Experimental versus In-Corpus Variation in Referring Expression Choice

**T. Mark Ellison, Fahime Same**
University of Cologne
{t.m.ellison,f.same}@uni-koeln.de

### Abstract

In this paper, we compare the results of three studies. The first explored feature-conditioned distributions of referring expression (RE) forms in the original corpus from which the contexts were taken. The second is a crowdsourcing study in which we asked participants to express entities within a pre-existing context, given fully specified referents. The third study replicates the crowdsourcing experiment using Large Language Models (LLMs). We evaluate how well the corpus itself can model the variation found when multiple informants (either human participants or LLMs) choose REs in the same contexts. We measure the similarity of the conditional distributions of form categories using the Jensen-Shannon Divergence metric and Description Length metric. We find that the experimental methodology introduces substantial noise, but by taking this noise into account, we can model the variation captured from the corpus and RE form choices made during experiments. Furthermore, we compared the three conditional distributions over the corpus, the human experimental results, and the GPT models. Against our expectations, the divergence is greatest between the corpus and the GPT model.

**Keywords:** referring expressions, variation, large language models, crowdsourcing experiment

## 1. Introduction

Linguistic communication conveys meanings with a range of different referring expressions (REs). The choice of RE depends on the discourse context in which they are realised, and how accessible they are to the speaker and addressee. The following examples refer to an imaginary character - Simon Brown, a famous portrait painter. *1. Simon Brown, the famous portrait painter, will attend the ceremony*, *2. Simon Brown will attend the ceremony*, and *3. He will attend the ceremony*. In sentence 1, the speaker supplies additional information about the referent's identity, i.e., that he is a famous portrait painter. Such expressions are used when the speaker presumes that the listener does not know the referent and would benefit from additional information, either to identify the referent or to add to their world knowledge. In contrast, sentence 2 uses a proper name without additional information. This RE type is appropriate when the speaker assumes that the listener needs only enough information to distinguish the referent from other potential referents. Finally, in sentence 3, the speaker uses a pronoun ("he") as an RE. This use occurs when the referent is established and salient in the context, and thus the addressee will know who is most likely to be referred to. These examples demonstrate the complexity of RE choice.

Reference production studies have categorised REs into various taxonomies of Referring Expression Forms (REFs) – including categories such as *pronoun*, *definite description* and *proper name* – and then sought to explain what conditions the choice of form. According to Accessibility Theory (Ariel, 2001), the more accessible a referent is, the more attenuated the corresponding REF. Other theoretical approaches are similar, linking the choice of REF to the salience, givenness, centrality, and/or discourse prominence of a referent (Gundel et al., 1993; Grosz et al., 1995; Chiarcos, 2011; von Heusinger and Schumacher, 2019). Linguistic studies have identified several factors that affect the choice of REF, including grammatical role, competition, thematic role, animacy, recency and coherence (Stevenson et al., 1994; Brennan, 1995; Arnold and Griffin, 2007; Kehler et al., 2008; Kaiser and Trueswell, 2011; Fukumura and van Gompel, 2011). For example, the *recency* factor measures the distance between a referent and its antecedent, with a recent antecedent usually facilitating a pronominal REF.

While some forms are found in some contexts more than others, RE choice is ultimately nondeterministic. Frequently, more than one form, and more than one expression, are equally acceptable to convey a reference. Castro Ferreira et al. (2016a) showed the non-deterministic nature of referring in an experiment in which multiple participants chose REs for the same referent in the same context. Castro Ferreira et al. compiled the results of their experiment as a corpus, known as *VaREG* (`https://ilk.uvt.nl/~tcastrof/vareg/`). The researchers provided participants with texts where references to the main topic had been replaced with gaps. The participants were tasked with filling the gaps with appropriate references. The experiment was balanced so that each RE slot was seen and filled by 20 participants.

Castro Ferreira et al. (2016a) classified the REs produced by humans into five classes of REF: *pronoun*, *proper name*, *description*, *demonstrative* and *empty reference*. They measured the entropy of the REF choices made by different participants for the same referential slots. The entropy varied considerably from one slot to another. They also investigated the impact of various linguistic factors such as recency, referential status and grammatical role on the relative entropy of the REFs, in order to assess their influence on REF variation. As an example, they observed that greater variation occurred in expressing the object in a transitive sentence than the subject.

Ellison and Same (2022) argue that although Castro Ferreira et al. (2016a)'s approach to investigating variation in REF choice is appealing, it is frequently prohibitive in terms of both time and cost. Instead, they propose to "infer variation in human behaviour through the variation found within a corpus of texts, gaining the benefits of understanding human variation without the substantial cost of human-intensive studies" (Ellison and Same, 2022, p. 2989). To capture *in-corpus* variation, they suggest identifying distinct classes of linguistic contexts made from the aggregation of various linguistically-informed feature-value combinations (henceforth called *feature-value categories*, *context categories*, or just *categories*). A possible category might encompass REs that feature a human referent in subject position and one sentence away from its antecedent (grammatical role: `subject`, animacy: `human`, recency: `one sentence away`). Ellison and Same (2022) adopted the working hypothesis that the same set of feature values will condition the same distribution over possible referring expression forms. Figure 1 provides an abstract scheme of the creation of such feature-value categories.

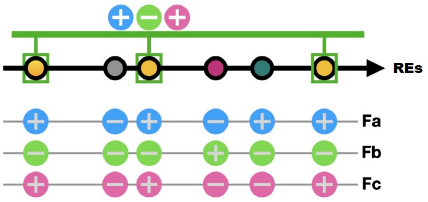


Figure 1: The initial row shows various REs within a corpus. The binary features Fa, Fb and Fc, characterise linguistic contexts, with categories formed from each distinct feature-value combination. For instance, the REs highlighted in yellow share the category defined by {+blue, -green, +pink}. From the subcorpus of all REs in a category, i.e. sharing the same feature-value combination, we can infer a distribution over REFs for that category, namely the relative frequency of forms in that category.

Since the VaREG corpus of Castro Ferreira et al. (2016a) is small (563 REs), Ellison and Same (2022) measured per-category distributions of REFs in the much larger Wall Street Journal (WSJ) portion of the OntoNotes corpus (Weischedel et al., 2013). They then compared these distributions against the human variation found within the VaREG corpus. Their comparison shows parallels in the entropy of the distributions from the WSJ to those of the human choices in VaREG.

However, this study has two significant drawbacks. Firstly, the authors provide only provisional similarities between the entropy patterns. Secondly, they compare their WSJ-inferred distributions with the human-made distributions derived from the VaREG corpus. The question arises as to whether similar entropy patterns would be observed if the in-corpus WSJ distribution were compared with results from a VaREG-like experiment with stimuli based on WSJ documents. The current study has responded to this question with a crowdsourcing experimental study of RE variation based on a subset of the WSJ corpus. The results of this study support an in-depth comparison of in-corpus variation and the variation obtained from parallel human choices. For ease of reference, we will call these experimental results Human, or our *experimental* results. In contrast, we will refer to the distributions of REF inferred from the WSJ as Corpus.

Since the advent of Large Language Models (LLMs), several studies have considered whether LLMs, e.g. GPT models, can simulate cognitive and linguistic human behaviours (Aher et al., 2023; Lampinen, 2022; Binz and Schulz, 2023; Dillion et al., 2023). Binz and Schulz (2023) demonstrated that systems can serve as accurate models of human cognitive behaviour in psychological experiments, e.g. about decision-making. Aher et al. (2023) introduced the term *Turing Experiment* for the reproduction of an existing experiment in psychology, economics or psycholinguistics, but using LLMs in place of human participants Their work is interesting as a way of understanding what human behaviours are captured by LLMs. Following these studies, we replicate our crowd-sourcing experiment using an LLM (GPT-4) in order to compare the REF choices made by human participants and by these models. We will refer to our corpus of machine responses Gpt.

This paper is organised as follows: section 2 states our research questions and hypotheses. In section 3, we present an overview of both the WSJ corpus and the corpus of variation collected in our crowdsourcing experiment. In addition, it describes our adaptation of the experiment to LLMs. The results are discussed in section 4, while section 5 offers an analysis of these findings and points to possible future extensions of this work.

# 2. Research Questions and Hypotheses

As mentioned, this paper presents a corpus of human variation, Human, based on the WSJ corpus. The parallel choice of REs by human participants lets us compare inter-personal variation in REF choice with in-corpus variation found in the WSJ itself (captured in Corpus), as well as variation in multiple runs of the same experiment in an LLM. In what follows, we give a detailed overview of the hypotheses that we test with this data.

**Feature-value combinations.** Ellison and Same (2022), drawing on substantial previous literature (Greenbacker and McCoy, 2009; Kibrik et al., 2016; von Heusinger and Schumacher, 2019), introduce a set of optimal features for predicting REFs over the WSJ corpus (Same and van Deemter, 2020). We use the same feature set to build the context categories, as depicted in figure 1. Because these features were selected as predictors of referring expression classes in human-generated text corpora, having the same values for these features results in the same distribution over REFs. We therefore hypothesise ($\mathcal{H}_1$) that *human variation in REFs is lower for REs belonging to the same category*, i.e. REs having the same values for all five features, rather than arbitrary pairs of REs. If confirmed, this hypothesis will imply that the features indeed capture (at least partially) the variation found in human REFs.

**In-corpus and human experimental variation.** One motivation for this research is to validate the use of category-conditioned, corpus-inferred distributions as models of variation in human RE choices. Such validation would require us to find similarity in the REF distributions in Corpus and those found in the experimental results, i.e. in Human. If the feature-defined categories and the corresponding inferred distributions truly reflect human variation, then for matching (as opposed to non-matching) categories, we should see the same distributions in all kinds of human-constructed REs. Thus our second hypothesis ($\mathcal{H}_2$) is that *the linguistic context-conditioned in-corpus variation and human experimental variation are more similar when the categories are matched than when they are not*.

**Experimental effects.** Creating REs to fill a slot in an online experiment is a radically different task to producing REs in a naturalistic setting, such as talking or writing an article. Consequently, we anticipate mismatches between the referential strategies employed in such experiments and those employed by authors, such as those writing the WSJ articles. One reason for these mismatches is that the experimental participants have much less context backing their selections than the authors of the articles. As a result, the participants in the experimental paradigm received special priming for one referent only, desensitising them to other referents in the text, and potential competition between them. With less sensitivity to potential competition, they are more likely to use pronouns, where a richer RE would be more suitable. Furthermore, as the experimental participants were not domain experts, producing informative definite descriptions would be more difficult for them as they lack in-depth knowledge of the referent. We hypothesise that if we see a difference between the experimental and in-corpus distributions, then *the experiment will show greater use of pronouns than the original corpus* ($\mathcal{H}_3$), and *there will be fewer definite descriptions in the REs produced by the participants than the original authors* ($\mathcal{H}_4$).

In addition to the inherent disparity between writing and gap filling employed in this study, a significant amount of noise (random selection of RE) will occur in the experiment, due to individual variation, the absence of an error-correcting editor, language skill level, etc. This noise may detrimentally affect the similarity between corpus-derived and experimental distributions over forms. We hypothesise that *human experimental data will show greater randomness than corpus data for the same context categories* ($\mathcal{H}_5$).

**Large Language Models.** We are also interested in finding out how well Large Language Models perform on the same task. For this reason, we replicated the human experiment using GPT-4 in place of actual humans. GPT-4 and other LLMs, owing to their vast training input, are able to bring more context to stimuli texts than non-expert experimental participants. Thus we expect the LLMs to behave more like the domain experts who wrote the original articles. Therefore, we hypothesise that the LLM-generated distributions will more closely align with the corpus distributions than with those produced by humans in experiments ($\mathcal{H}_6$). Continuing from ($\mathcal{H}_4$), a second expected consequence of greater domain knowledge is that the number of descriptions produced by the GPT-4 models will exceed those produced by human participants ($\mathcal{H}_7$). We would expect to see more descriptions in Corpus than Human ($\mathcal{H}_4$), and so expect Gpt to also follow this pattern ($\mathcal{H}_7$). In section 4 we explore the extent to which these hypotheses are confirmed by our data.

# 3. Experiment and Analysis

This section begins with the corpus and the set of features used in the study (3.1). We then outline the experiments (Human and Gpt) using stimuli constructed from the corpus (3.2).

## 3.1. The corpus

The WSJ corpus consists primarily of financial news articles. We used a total of 30,439 REs extracted from the WSJ to explore the distribution of REFs. This study focusses on three forms: pronouns, proper names and descriptions. The global distribution of these forms in the corpus is: 12 005 (39.44%) descriptions, 11 153 (36.64%) proper names, and 7 281 (23.92%) pronouns.

We use the five features described in Ellison and Same (2022) to construct feature-value categories. The distributions for each category are constructed, and these together form the model CORPUS. The features and their possible values are: `Grammatical Role` (subject, object, possessive determiner), `Form of the antecedent` (pronoun, proper name, description, or first-mention of the referent in the text), `Animacy` (human, other), `Sentence recency`, i.e., the recency of the referential antecedent measured in sentences (same sentence, different sentence, first-mention), and `Paragraph recency` (same paragraph, different paragraph, first-mention).

## 3.2. The crowd-sourcing experiment: data collection

**Material.** To ensure that we have a representative sample of RE uses, we defined four categories of referents as follows: human, city or country, organisation, other (including concrete objects or abstract concepts). We only used documents in which the topic has at least seven mentions and chose texts with the following distributions:

| Referent type | Documents (n) |
|---|---|
| Human | 20 |
| Organization | 10 |
| City & country | 10 |
| Other | 10 |

Table 1: Number of texts selected according to the referent type.

**Participants.** 100 Participants were recruited through Amazon Mechanical Turk (MTurk). To ensure consistency, we restricted MTurk workers to those located in the United States, with an approval rating of $\geq$ 95% and 1000 or more HITs approved. Among the participants, 56 were male, 39 were female, and 1 identified as other. The average age of the participants was 38.6 years old (ranging from 24 to 73). A majority of the participants (83 individuals) were native English speakers.

**Experiment design and procedure.** The experiment consisted of 50 texts divided into 5 lists. Each list contained four documents with a main human referent, two documents with a city or country, two documents with an organisation, and two documents categorised as "other".

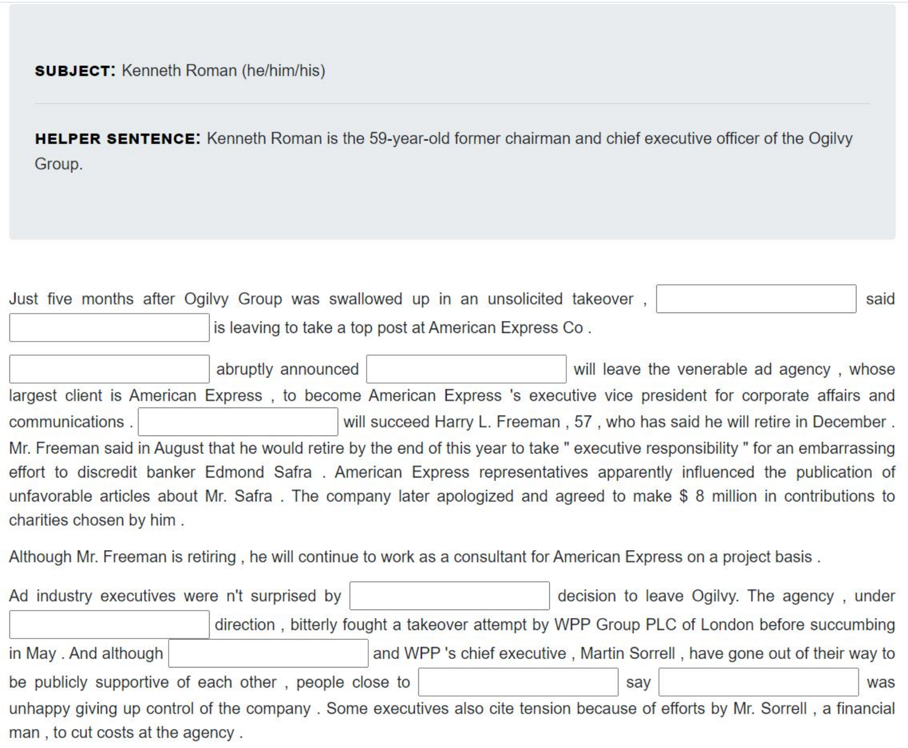
Prior to the experiment, participants received an introduction outlining the procedure and requesting their consent to participate. At this point, they were asked to provide information about their age, gender and proficiency in English.

Participants were presented with one of the lists containing ten documents. Mentions of the article topic were replaced with gaps. Participants were instructed to fill in each gap with an RE referring to the given referent. To familiarise participants with the referents, a full RE representing the referent was selected as the subject of the text and pronominal forms were given in parenthesis. To assist participants in generating more informed and descriptive REs, where they might choose to do so, a helper sentence was provided, offering background information about the referent. One of the experimental items is shown in figure 2.

**Annotation.** The experiment involved 414 referential gaps, comprising 31.3% pronouns, 42.1% proper names and 26.6% descriptions. The participants produced a total of 8 280 REs. We annotated the REFs of these REs with three categories, namely pronoun (e.g., *he*), proper name (e.g., *Kenneth Roman*) and description (e.g., *the country*). Of these REs, 3 022 (36.5%) were annotated as proper names, 3 484 (42.1%) as pronouns, 1 063 (12.8%) as descriptions, and the remaining 710 (8.6%) cases were classified as *unacceptable*. These were cases where the participants' responses could not be categorised as belonging to either of the three REFs. The majority of these cases were those that did not refer at all or did not refer to the target entity.

## 3.3. The GPT Experiment

To conduct the LLM experiment, constructing GPT, we used OpenAI's GPT-4 (model=gpt-4). The prompt included the same instructions that were given to human participants, with only minor modifications to adapt these for the language model. To maintain fidelity to the experimental setting used in the human experiment, we used the same lists of items. Each list was run 20 times. Separate connections were used each time to eliminate the chance of the results being confounded by the previous runs. As in the human experiment, we annotated the generated expressions for their referential forms. To accomplish this, we used GPT-4 once more, this time to annotate two specific pieces of information: (1) the form of the RE, and (2) whether the generated RE accurately referred to the intended referent. After this step, all annota-



**SUBJECT:** Kenneth Roman (he/him/his)

**HELPER SENTENCE:** Kenneth Roman is the 59-year-old former chairman and chief executive officer of the Ogilvy Group.

Just five months after Ogilvy Group was swallowed up in an unsolicited takeover , □ said □ is leaving to take a top post at American Express Co .

□ abruptly announced □ will leave the venerable ad agency , whose largest client is American Express , to become American Express 's executive vice president for corporate affairs and communications . □ will succeed Harry L. Freeman , 57 , who has said he will retire in December . Mr. Freeman said in August that he would retire by the end of this year to take " executive responsibility " for an embarrassing effort to discredit banker Edmond Safra . American Express representatives apparently influenced the publication of unfavorable articles about Mr. Safra . The company later apologized and agreed to make $ 8 million in contributions to charities chosen by him .

Although Mr. Freeman is retiring , he will continue to work as a consultant for American Express on a project basis .

Ad industry executives were n't surprised by □ decision to leave Ogilvy. The agency , under □ direction , bitterly fought a takeover attempt by WPP Group PLC of London before succumbing in May . And although □ and WPP 's chief executive , Martin Sorrell , have gone out of their way to be publicly supportive of each other , people close to □ say □ was unhappy giving up control of the company . Some executives also cite tension because of efforts by Mr. Sorrell , a financial man , to cut costs at the agency .

Figure 2: Example of an item from the crowdsourcing experiment. Participants were tasked with supplying referring expressions at each slot for the referent shown at the top of the item (the *subject*) and described in more detail in the *helper sentence*.

tions were manually reviewed for potential errors and inconsistencies.

## 4. Results Relative to the Hypotheses

This section compares the experimental results in Human and Gpt with the distributions in Corpus and explores the relationships between them. The discussion is focussed on finding support, or otherwise, for our hypotheses from section 2. Since uninterpretable responses from participants do not bear on our hypotheses, we exclude 'unacceptable' cases from this analysis.

$\mathcal{H}_1$ **variation patterns are more similar within categories than between them.** In Human, 414 REs from the WSJ became slots for participants to fill. Each was filled by 20 participants. However, a number of the 8 280 responses had to be discarded as uninterpretable. The 20 REs collected for each of these slots form a distribution over answer forms. If we identify slots by labels $s$, we can write $n_s^f$ for the number of times a REF $f$ occurs in the answers provided for this slot. Note that throughout this paper, as in tensor notation, superscripts are indices rather than powers. Each of these occurrence counts can be recast as a relative frequency distribution $r_s^f = n_s^f / \sum_g n_s^g$.

Like Castro Ferreira et al. (2016a), Castro Ferreira et al. (2016b) and Ellison and Same (2022), we compare distributions of variation using the Jensen-Shannon Divergence (JSD) measure. This symmetric metric measures how much two distributions differ from each other. It sums the asymmetrical $KL$ (Kullback-Liebler) divergence measure of each distribution from the average of the two. The definition of the $JSD$ and $KL$ measures are shown in equations (1) and (2) respectively. In these definitions, $\mathbf{d} = (d^f)_{f\in F}$ is a distribution over a set $F$ of values $f$. Two different distributions are notated by $\mathbf{d}_1$ and $\mathbf{d}_2$.

$$\begin{aligned} JSD(\mathbf{d}_1, \mathbf{d}_2) &= \frac{KL(\mathbf{d}_1||\mathbf{d}_{12}) + KL(\mathbf{d}_2||\mathbf{d}_{12})}{2} \quad (1) \\ \text{where } \mathbf{d}_{12} &= \frac{\mathbf{d}_1 + \mathbf{d}_2}{2} \\ KL(\mathbf{d}_2||\mathbf{d}_1) &= \sum_{f\in F} d^f \log \frac{d_2^f}{d_1^f} \quad (2) \end{aligned}$$

For each pair of unique REs, $s_1$ and $s_2$, that occur in Human, we look at the Jensen-Shannon divergence (JSD) score between the distributions returned by the participants for those REs, i.e. $JSD(\mathbf{r}_{s_1}, \mathbf{r}_{s_2})$. Note that we are here comparing only distributions take from Human with each other, not with distributions from other data sets, i.e. Corpus or Gpt. For ease of visualisation, we take all logarithms here to be base 2.

Figure 3 shows two density plots structured as a split violin graph. The green density plot on the left shows the distribution of $JSD$s for all pairs of slots in HUMAN. The right, orange density plot shows the distribution of $JSD$s where the slots being compared belong to the same feature-defined category. The figure shows a strong reduction in $JSD$s when categories match, with the median $JSD$ across all pairs being greater than the 3rd quartile value from the matching pairs. This result supports hypothesis $\mathcal{H}_1$.

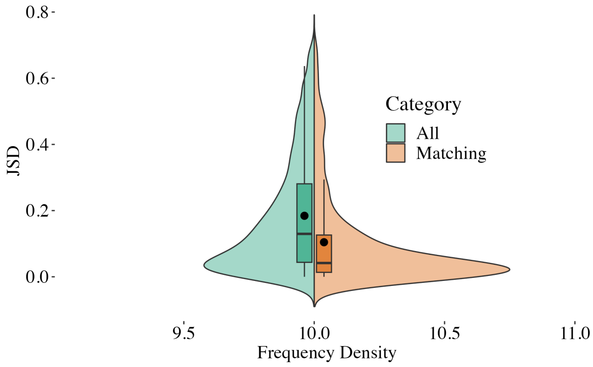


Figure 3: Intra-class vs. between-class JSD scores. In the box plots, the dots show the means. The boundaries of the box and the centre line show the quartile points and the median respectively. The whiskers mark out 95% of the data.

**Description Length.** In order to see how strong support is for hypothesis $\mathcal{H}_1$, we look at the encoding length of the experimental data relative to matching categories. This is then compared to the average encoding length obtained from random alignments of categories. Intuitively, the encoding length is the smallest number of bits needed to losslessly encode the model and the data together (see e.g. Rissanen, 1978,Wallace and Dowe, 1999,Grünwald, 2007). This encoding length, denoted by $EL(n^f|p^f)$, combines the count $n^f$ of forms of a type $f$, given their optimal encoding, based on their probability $p^f$. It is defined in (3).

$$EL(\mathbf{n}|\mathbf{p}) \quad = \quad -\sum_f n^f \log p^f \tag{3}$$

For example, suppose we have 10 pronouns, 6 proper names and 4 descriptions chosen by participants to fill a particular slot. The probabilities $p^f$ of these REFs are respectively 0.5, 0.3 and 0.2, while their counts $n^f$ are 10, 6 and 4. The total information needed to encode these results are therefore $-10\log_2 0.5 - 6\log_2 0.3 - 4\log_2 0.2 = 29.71$ bits.

The bit-length difference in two encodings is the negative logarithm of the Bayes' Factor comparison of the two defining models. In other words, shorter encodings are provided by better models, and much shorter encodings by much better models.

The optimal encoding model offers a different encoding of REFs for each RE slot. However, this is a false economy, as it requires many bits to encode the distribution of responses separately for each RE slot, information that is needed to fix the per-slot encodings. This cost is the number of independent probabilities which need to be specified, multiplied by the (possibly fractional) number of bits needed to express each. We use $N$ equally-spaced buckets on the unit interval to discretise probabilities for finite representation (see figure 4). Any of the probabilities to be specified is mapped onto its containing bucket $\left[\frac{i-1}{N}, \frac{i}{N}\right)$. The probability used in practice is then the centre point of the bucket, normalised against the other members of its distribution.

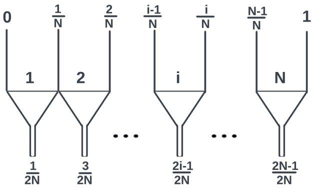


Figure 4: Finite precision representation of probabilities in $N$ buckets on the unit interval. If the value to be represented is $p$, then the finite precision version will be $\frac{\lfloor pN \rfloor + \lceil pN \rceil}{2}$.

Table 2 shows the fraction of information in the experimental results which is accounted for by the model. The baseline account tries all distributions available in the model class to encode the experimental data and averages the encoding lengths. A maximum entropy model of the data (equal probability encodings of description, name, and pronoun) is taken as the floor in performance. For our dataset, this quantity was: 11 997 bits. The ceiling in performance is obtained when we model the distributions of REFs by category in terms of the frequencies gained in the experiment itself, to a precision of 7 bits, giving a total of 9 031 bits for representing the experimental results. We take this as our 100% success value.

The graph in figure 3 corresponds to the model *RE E* in table 2. The encoding length for the *All* case is greater than that for the *Matching* case by 7592 bits, corresponding to a Bayes' factor of approximately $10^{2285}$. We can say that the difference in the density plots of 2 is statistically significant. Thus matching categories do result in significantly less unaccounted variation in HUMAN - reflected in the shorter encoding - than when categories are not matching.

$\mathcal{H}_2$ **in-corpus and experimental variation are more similar when aligned.** This hypothesis was tested by comparing the JSDs of distributions of

| Model | N | Matching (b) | Matching(%) | All (b) |
|---|---|---|---|---|
| MaxEnt | | 11 996.6 | 0.0% | |
| RE E | 21 | 10 636.3 | 45.9% | 18 228.4 |
| Cat E | 7 | 9 030.9 | 100.0% | 13 706.6 |
| Cat C | 3 | 10 802.5 | 34.1% | 14 192.8 |
| Diff C | 3 | 9 677.2 | 78.2% | 11 697.6 |

Table 2: Description length of Human encoded by different models. The columns give the model name, the precision (dividing probability space into N parts), the number of bits needed to encode the corpus with matched IDs, and the average number of bits required across arbitrary matchings of the RE or category IDs. The models are: **MaxEnt** equal length fixed encodings for all three REFs; **RE E** a new, precise encoding is defined for each RE slot, with probabilities from frequency in the experimental results themselves; **Cat E** a new encoding is defined per category with probabilities from Human; **Cat C** a new encoding is defined per category with probabilities from Corpus; and **DiffC** applies the diffusion matrix to **Cat C**.

REFs found for each category in Corpus and in Human. First, for each category, we compute the relative frequency of REFs from that category. Subsequently, we repeat the process in Human, looking at the distribution over responses for all slots from each category. For example, there are 661 cases in the corpus where the REs referring to *humans* occur in the *subject* position within the *same paragraph* but in a *different sentence* from their corresponding *pronominal* antecedent. Of these 661 cases, 458 (69.29%) are pronouns, 147 are proper names (22.24%), and 56 (8.47%) are descriptions. Examining the same feature-value category in the experimental results, there are 369 cases (excluding unacceptables). Among these, 203 (55.01%) are pronouns, 137 (37.13%) are names, and 29 (7.86%) are descriptions. The JSD of these two distributions is $0.0196$.

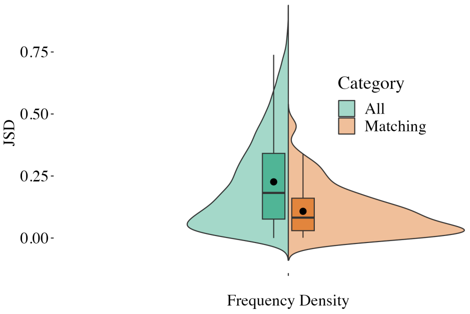


Figure 5: JSDs of corpus and experimental form distributions with identical (Matching) and arbitrary (All) feature-value combinations.

The JSDs comparing Human and Corpus when categories are matching or associated arbitrarily are shown in figure 5. Once again, we see that the median JSD for the *Matching* spread is less than the lower quartile of the *All* spread, suggesting a strong effect of category alignment.

The corresponding line in table 2 is **Cat C**. Here we see that there is a difference in representational length of 3 390.3 bits between category-aligned and arbitrarily matched encodings. This encoding length difference reflects a Bayes' Factor of the order of $10^{1020}$. The model with matching categories is substantially superior to those with arbitrarily matched slots.

$\mathcal{H}_3$ **more pronouns in the experimental results,** $\mathcal{H}_4$ **fewer definite descriptions in the experimental results.** The raw values can be seen in figure 6. It is apparent that in the experimental results, we see fewer descriptions and more pronouns.

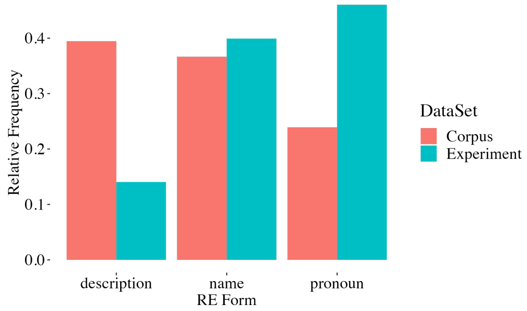


Figure 6: Relative frequencies of different forms in corpus and experimental results.

Because the results in Human were gathered on the basis of categories in Corpus, we can directly relate the REFs in the corpus to conditional distributions of REFs in the experimental results. These are visualised as a heatmap in figure 7.

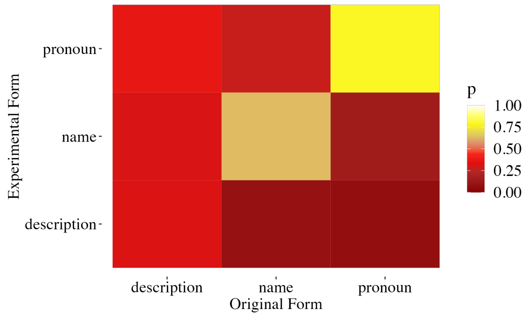


Figure 7: The conditional probability of experimental forms given the original form used in the corpus.

The heatmap shows a strong correlation between the pronominal realisation in Corpus and in Human. The central column of the heatmap, where the original corpus RE was a proper name, also exhibits substantial agreement. However, where there is a definite description in the corpus, these are not realised consistently in the experiment. They seem equally likely to be realised as ‘descriptions’, ‘proper names’ or ‘pronouns’. So $\mathcal{H}_3$ and $\mathcal{H}_4$ are tied. The lack of realisation of corpus ‘descriptions’ as ‘descriptions’ in the experimental results accounts for their fall in numbers in Human, and their redistribution accounts for the increase in pronominal forms.

$\mathcal{H}_5$ **human experimental participants produce noisier, and thus flatter, distributions of referring expressions.** One potential problem with using the relative frequencies from Corpus as a model of Human is that it overfits. Here we explore one way of relaxing this tightness of fit with Corpus: we add randomness that brings the overall distribution of REFs in Corpus into line with the distribution of REFs found in Human. We write $h^f_{f'}$ for the ‘diffusion‘ matrix of relative frequencies of $f$ in Human in slots where the corpus had REF $f'$. We can mimic the noise in the relationship between context categories and REFs in the following way. In each context $k$, where the distribution of REFs in the corpus is $q^{f'}_k$, we spread this distribution by multiplying it by the diffusion matrix $h^f_{f'}$. The resulting matrix of conditional probabilities $v^f_k = \sum_{f'} h^f_{f'} q^{f'}_k$ gives a new distribution for each category. As a result, it offers predictions more in line with the responses seen in Human. We say that the resulting conditional distribution is *diffused*.

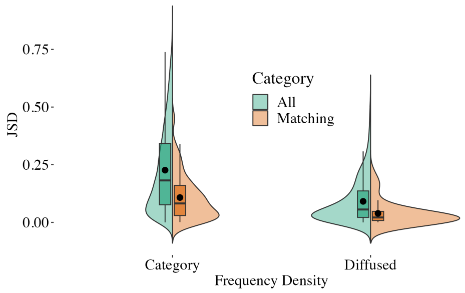


Figure 8: JSDs of the corpus and experiment distributions in non-diffused (Category) and diffused models.

We see how the non-diffused and diffused distributions match the experiment results in figure 8. The diffused model is a substantially better match for the experimental results, as is visible in its lower JSD values, indicating a better match between the distributions found in Human and Corpus (after diffusion).

The diffused conditional probabilities are better predictors of the data in Human than the non-diffused corpus relative frequencies. Even incorporating the cost of representing the matrix in bits of information (presuming a probability resolution of 0.002%), we find in table 2 that the diffused model **Diff C** outperforms all models, other than **Cat E**. Recall that **Cat E** offers the best possible encoding of Human because it uses distributions found there to predict, and so encode, its own data. **Diff C** has removed 78% of the redundancy in the maximum entropy model relative to **Cat E**.

$\mathcal{H}_6$ **proposed that Gpt REF distributions would be more similar to those of Corpus than to those of Human.** This hypothesis is not supported by the experimental results. Instead, as shown in figure 9, the mean JSDs found between categories in Gpt and Corpus are substantially higher than those between Gpt and Human. These results suggest that the LLM models are better accounts of the variation seen in Human, i.e. responses given by experimental participants, than the variation seen in Corpus, i.e. the thought-over and edited use of REs in newspaper articles.

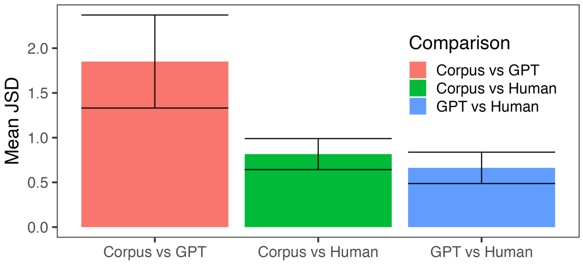


Figure 9: Mean JSDs comparing REF distributions in Gpt with those based on Corpus predictions, and also the distributions produced by human participants. As shown by the non-overlapping standard-error error-bars, there is a significant difference between the mean JSDs of Gpt and Corpus on the one hand (red), and Gpt and Human on the other (blue). This indicates that the Gpt distributions are not more closely aligned with the Corpus distrbutions, but rather with the Human ones, contradicting $\mathcal{H}_6$. Note that Corpus also shows low JSDs with Human, offering a model of experimental variation at a similar level to the LLM.

$\mathcal{H}_7$ **proposed that GPT would produce more descriptive REs than human participants**. This is indeed confirmed, as shown in figure 10. The difference in relative frequencies is substantial (34%: 0.188 for GPT, 0.140 for the human participants), confirming the hypothesis.

## 5. Discussion

The results in section 4 show positive evaluations of our hypotheses, except for hypothesis $\mathcal{H}_6$. In Human, the distributions of variation for REFs drawn from the same category are more similar than distributions from different categories. Comparing Human and Corpus, distributions over forms conditioned by category are more similar for matching

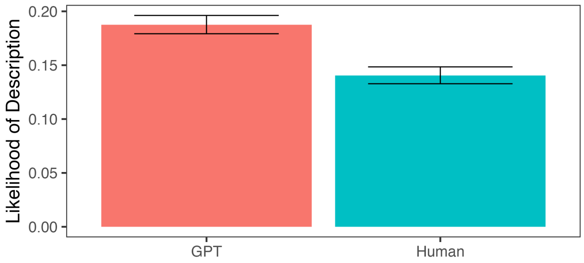


Figure 10: The likelihood of description (vs pronoun or name) use in experimental results: GPT vs HUMAN. The error bars show the Beta-Bernoulli 95% confidence intervals, and do not overlap, so the difference in results is unlikely to be due to chance.

categories than for non-matching categories. The pattern of forms in HUMAN reflects our prediction that there are fewer 'descriptions' and more 'pronouns'. Using a diffusion matrix to render the CORPUS conditional distributions more like those from HUMAN resulted in substantially better predictions of the experimental data.

Consideration of the variation in GPT and HUMAN led to two more hypotheses. Our hypothesis that the LLM would behave more like CORPUS than HUMAN proved incorrect. However, it was the case that GPT showed more use of descriptions than HUMAN.

Our aim in this paper has been to explore how well the variation seen in a large corpus, like WSJ, can function as a proxy for variation found in experiments. Both corpus and experiment are opportunities to explore how contextual feature-value categories condition inherent variability in REFs. This is reflected in the results in section 4, in the evaluation of hypothesis $\mathcal{H}_2$. Distributions defined by the same feature-value categories are much more similar than those defined by arbitrary matching. So there is evidence of a common cause at play in conditioning these distributions.

One interesting finding is the effect of overfitting. While the categorisations themselves allow good compression of the data (see table 2, model **Cat E**), the distributions constructed within the corpus offer only partial improvement over the flat distribution (34% at best). However, applying a mask mapping from corpus to experimental forms, derived from how the original forms in corpus slots are realised within HUMAN, resulted in superior performance, eliminating 78% of available redundancy. This mask has the effect of reconciling the difference in the overall distribution of REFs in the two data sets CORPUS and HUMAN. It also tends to flatten distributions, reducing the amount of overfitting, and so leading to a more general model of referential variation.

We compared the distributions of RE types from CORPUS with GPT-4's interpolation of referential forms (GPT), when it was assigned the same task as human experimental participants.

Comparing the three conditional distributions over the data sets, using the JSD, gives us three distance measures: CORPUS-GPT, CORPUS-HUMAN and HUMAN-GPT. The CORPUS-GPT divergence is the largest. While the HUMAN-GPT divergence is the smallest. These results together imply that both GPT and CORPUS are capturing substantial aspects of the variation found in human experimental variation. However, the fact that the variation in GPT and CORPUS are so different shows that they are not capturing the same aspects of human variation.

This result, in turn, suggests that corpus study can teach us things about referential type use not yet captured in LLMs. This conclusion also quietens concerns that GPT-4 might just reproduce forms from the corpus which may well have formed part of its training set.

The results presented here show that there is a substantial connection between variation we see in corpora and variation found in and between human participants in a variationist experiment. If we take care not to overfit, we can find substantial agreement between the two. Any remaining mismatch may result from differences in cognitive aspects of the language-production situation.

Possible future work could enrich these conclusions by looking at the lexical content of the referring expressions themselves, not just at their type. All associated data including the human and GPT-4 experimental results are publicly available and can be accessed on our GitHub repository: https://github.com/fsame/WSJ-VariationCorpus.

***Acknowledgements:*** We thank anonymous reviewers for their helpful comments. Both authors are supported by the German Research Foundation (DFG) – Project-ID 281511265 – SFB 1252 "Prominence in Language".

## 6. Bibliographical References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

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.

Mira Ariel. 2001. Accessibility theory: An overview. In Ted Sanders, Joost Schilperoord, and Wilbert Spooren, editors, *Text Representation: Linguistic and psycholinguistic aspects*, volume 8, page 29. John Benjamins Publishing Company.

Jennifer E. Arnold. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, 31(2):137–162.

Jennifer E. Arnold and Zenzi M. Griffin. 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, 56(4):521–536.

Marcel Binz and Eric Schulz. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Susan E. Brennan. 1995. Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 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*, pages 423–427, San Diego, California. Association for Computational Linguistics.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 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)*, pages 568–577, Berlin, Germany. Association for Computational Linguistics.

Christian Chiarcos. 2011. The mental salience framework: Context-adequate generation of referring expressions. In *Salience*, pages 105–140. DE GRUYTER MOUTON.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

T. Mark Ellison and Fahime Same. 2022. Constructing distributions of variation in referring expression type from corpora for model evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2989–2997, Marseille, France. European Language Resources Association.

Kumiko Fukumura and Roger P. G. van Gompel. 2011. The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, 26(10):1472–1504.

Charles F Greenbacker and Kathleen F McCoy. 2009. Feature selection for reference generation as informed by psycholinguistic research. In *Proceedings of the CogSci 2009 Workshop on Production of Referring Expressions (PRE-Cogsci 2009)*.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Peter D Grünwald. 2007. *The minimum description length principle*. MIT press.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

Elsi Kaiser and John C. Trueswell. 2011. Investigating the interpretation of pronouns and demonstratives in finnish. In *The Processing and Acquisition of Reference*, pages 323–354. The MIT Press.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L Elman. 2008. Coherence and coreference revisited. *Journal of semantics*, 25(1):1–44.

Andrej A Kibrik, Mariya V Khudyakova, Grigory B Dobrov, Anastasia Linnik, and Dmitrij A Zalmanov. 2016. Referential choice: Predictability and its limits. *Frontiers in Psychology*, 7:1429.

Andrew Kyle Lampinen. 2022. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *arXiv preprint arXiv:2210.15303*.

Massimo Poesio. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 154–162, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Fahime Same and Kees van Deemter. 2020. A linguistic perspective on reference: Choosing a feature set for generating referring expressions in context. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4575–4586, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rosemary J. Stevenson, Rosalind A. Crawley, and David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.

Klaus von Heusinger and Petra B. Schumacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117–127.

C. S. Wallace and D. L. Dowe. 1999. Minimum Message Length and Kolmogorov Complexity. *The Computer Journal*, 42(4):270–283.

## 7. Language Resource References

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. 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*, pages 423–427, San Diego, California. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0.