LLMs Don't "Do Things with Words" but Their Lack of Illocution Can Inform the Study of Human Discourse

Zachary P Rosen
University of California, Los Angeles
z.p.rosen@ucla.edu

Rick Dale
Professor of Communication
UCLA
rdale@ucla.edu

Abstract

Despite the long-standing theoretical importance of the concept of illocutionary force in communication (Austin, 1975), quantitative measurement of it has remained elusive. The following study seeks to measure the influence of illocutionary force on the degree to which subreddit community members maintain the concepts and ideas of previous community members' comments when they reply to each other's content. We leverage an information-theoretic framework implementing a measurement of linguistic convergence to capture how much of a previous comment can be recovered from its replies. To show the effect of illocutionary force, we then ask a large language model (LLM) to write a reply to the same previous comment as though it were a member of that subreddit community. Because LLMs inherently lack illocutionary intent but produce plausible utterances, they can function as a useful control to test the contribution of illocutionary intent and the effect it may have on the language in human-generated comments. We find that LLMs indeed have statistically significant, lower entropy with prior comments than human replies to the same comments. While this says very little about LLMs on the basis of how they are trained, this difference offers a quantitative baseline to assess the effect of illocutionary force on the flow of information in online discourse.

Introduction

Central to all speech acts is the speaker's intentionality, or illocutionary intent (Austin, 1975), encoded in their utterances. Even the most seemingly inoccuous utterances can be, and often are, laden with unspoken intent. Indeed, consider the following example: "We're out of milk." While the surface form can easily be read as a statement of fact, one is hard pressed not to look for additional intent lurking just under the surface – in this case, the intention to head to the store or make a request of a roommate. Illocutionary intent serves as a primary motivation for communication in general. Despite this importance, it has been fundamentally difficult to measure the degree to which it affects the variability seen in communicative acts.

The following report describes an experiment designed to measure the contribution of illocutionary intent by combining information theory, convergence modeling, and large language models as proxies for illocution-deficient language generators. In the following sections we will first frame our study in terms of the existing literature on pragmatics and speech act theory. We will then discuss the utility of large language models (LLMs) as exemplars of linguistic behavior without human intentionality. We will then describe our experimental design and showcase results. These results quantify how human communicators *diverge* from LLMs under the illocutionary force they deploy in online contexts. We conclude with a discussion of these findings and their implications for the study of human communication and cognition.

Language as action

Central to contemporary work on Speech Act Theory is J.L. Austin's division of utterances into what are effectively three parts or, as he dubbed them, forces: the locutionary force (the surface or truth conditional meaning of an utterance), the illocutionary force (the implied meaning or action that the speaker wants to elicit from the listener), and the perlocutionary force (the change effected in the environment through the use of that particular speech act) (Austin, 1975). A number of refinements to Austin's work have focused on the necessity of illocutionary intent in communication as a central tenet of their own work (Sperber & Wilson, 2001; Heintz & Scott-Phillips, 2023; Holtgraves & Ashley, 2001; Enfield & Sidnell, 2022; Oller & Griebel, 2021). For example, Enfield and Sidnell (2022) point out that "language and human *intersubjectivity*," which is understood as the expected actions that interlocutors should take when engaging with one another in conversation, "are co-constituting." This sentiment is mirrored in a number of other research programs, such as those which emphasize that the interpretation of meaning is "driven by the immediate communicative challenges of daily life", and further highlights how need, and thus intention, drive collaborative understanding (Christiansen & Chater, 2022). In fact, it may be that language itself evolved in order to communicate complex intentions between agents (Heintz & Scott-Phillips, 2023).

LLMs: Language sans illocution

Now that we've established the particular speech act phenomena we want to assess, let's turn our attention to describing what LLMs are and how they might be useful in measuring illocutionary intent.

LLMs are supersized versions of a language modeling technique introduced in 2018 by researchers at Google (Brown et al., 2020) – the transformer language model (Vaswani et al., 2017; Devlin, Chang, Lee, & Toutanova, 2019). The model itself allows researchers to capture extremely subtle distinctions in word usage across various contexts, significantly advancing long-standing approaches in

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

vector mechanisms for representing meaning (Landauer, McNamara, Dennis, & Kintsch, 2013; Pennington, Socher, & Manning, 2014; Mikolov, Chen, Corrado, & Dean, 2013). Contemporary LLMs are direct descendants of the original Generalized Pretrained Transformer (GPT), which is a transformer language model trained on next token prediction rather than a cloze task (Radford et al., 2019). With their billions of parameters to earlier transformers' million of parameters, they have improved memory capacity for storing patterns in language as encountered in their training data (Frankle & Carbin, 2019).

A number of claims have been made about emergent properties arising from the massive increase in size and data reflected in LLMs. These include the potential to mimic human performance in a number of psycholinguistic tasks (Dillion, Tandon, Gu, & Gray, 2023; Trott, Jones, Chang, Michaelov, & Bergen, 2023; Cai, Haslett, Duan, Wang, & Pickering, 2023), understanding cognitive states (Trott et al., 2023), and being capable of generating passable (if bland) synthetic discourse (Byun, Vasicek, & Seppi, 2023). This performance is likely based on knowledge of statistical distributions of word usage, as has been suggested in prior work (Futrell & Hahn, 2022). Such performance can still prompt familiar debate in philosophy of mind about whether the underlying processes involved in that performance are properly "cognitive" in the way we would infer in the human case (Harnad, 1990).

Despite such debate, one key human linguistic factor that appears to be lacking in LLMs is illocutionary intent. Understanding the crux of this statement requires some knowledge of how LLMs are trained, and the difference in that process and human language acquisition with regards to the role of intentionality.

LLMs require initial training on massive corpora of text data taken from the internet. That training process consists of a simple next token prediction task, wherein a model is presented with some amount of prior context and then predicts what the next word in that sequence might be. After a guess has been made, it is compared to the actual next word in the sequence as represented in the original text, and the model's internal weights are updated according to how "off" its guess was (Brown et al., 2020; Mikolov et al., 2013; Devlin et al., 2019; Pennington et al., 2014). More recently LLM training is augmented with several rounds of "Reinforcement Learning with Human Feedback" (RLHF), where the model's outputs are then shown to a human annotator and that annotator gives an error signal to the model depending on whether the annotator thought that the model response was appropriate or not (Stiennon et al., 2020).

While RLHF introduces an additional source of complexity in training – the need for the LLM to produce text that both plausibly continues from the input it received from a user *and* is subjectively acceptable to that same user – the task the LLM is trained on is still simply next token prediction (Brown et al., 2020) though tuned with a joint probability on next tokens based on what is acceptable to the sensibilities of the human giving feedback.

Readers familiar with these statistical approaches may point out that this process resembles next word prediction in human subjects for language comprehension tasks, and a number of works would seem to support this theoretical claim (Futrell & Hahn, 2022; Levy, 2008). However, it could be argued that there is a clear difference in how LLMs encode transitional probabilities for the purpose of sequence prediction versus the motivations underlying language acquisition in children. When children learn language, the initial objective isn't simply to predict the next word correctly, but ultimately to elicit some preferable change in their environment (Oller & Griebel, 2021). Indeed, much of children's early language interactions appear to center around modulating the language of their requests in order to maximize the likelihood of reaching some desired state of affairs or intersubjective understanding (Clark, 1974) – a task that is often accomplished by recycling bits of language heard in prior interactions with caregivers, and filtered through children's own prior language use (Snow & Goldfield, 1983). As adults, this link between linguistic expression and environmental control remains at the heart of language use. A number of contemporary theories center peoples' intersubjective understanding that language functionally expresses individuals' internal mental state, and that said mental state is linked to the needs and desires of interlocutors (Christiansen & Chater, 2022; Enfield & Sidnell, 2022; Heintz & Scott-Phillips, 2023; Oller & Griebel, 2021).

While next word prediction may help LLMs predict transitional probabilities between likely continuations of strings of tokens, there is no reason to believe that next word prediction alone would necessarily allow for the emergence of *intentional* language use (Zarcone, van Schijndel, Vogels, & Demberg, 2016; Aurnhammer & Frank, 2019; Sperber & Wilson, 2001). Instead, if our current understanding of human language acquisition holds true, illocutionary intent precedes next word prediction. And while the success of LLMs in generating plausible text, however impressively, shows that it is not necessary to have an internal sense of "need" in order to acquire lexical and syntactic mastery of a language, there is nothing in an LLM's training process that indicates that they are learning to use language intentionally rather than writing a plausible continuation to whatever preexisting intentions or wants were expressed by the humans that interact with them – at least, not when we compare the development of linguistic competency in humans versus LLMs. And while a skeptic can certainly still claim that this does not preclude the emergence of intentionality, it is necessary for such a claim to be presented in conjunction with extraordinary evidence in order to reject the null hypothesis that next word prediction is insufficient to acquire intention.

Experiment and Design

Data

Our data collection approach can be divided into two separate processes. First, we collected example replies to Reddit sub-

missions. Second, we then recorded responses for ChatGPT to those same comments which had also been replied to by other Redditors online.

To collect our Reddit data, we used the PRAW Python API for querying the website. We focused our analysis on the following subreddit communities r/StarWars, r/StarTrek and r/Dogs. These communities were selected in order to limit the potential for politically or emotionally charged content being analyzed. We then indexed the comment id, parent id (i.e., if the comment was a reply, which previous comment was it replying to) and submission ids for the 100 most recent submissions in each subreddit up until October 29th, 2023. We then found all comments within those posts that had at least 5 replies and retained only those comments and the associated replies for our data set. We then went back to this index, converted the comment text for the comments in our index to word vectors using RoBERTa-base-uncased provided free and open source from the Hugging Face, Inc. library (Wolf et al., 2020), and also collected the total number of upvotes (also called "karma") for the comments in our index for analysis. We also collected an anonymized id for author names. This was done to ensure that we could not compare comments or replies written by the same author to each other. In an effort to maintain Redditors' privacy, we will not make any text or identifying information available to other researchers, but can share our indexes upon review. No model was fine-tuned or explicitly trained on the data we analyzed.

In total, this yielded a dataset of approximately 179 total comments for analysis, and 4,045 total replies.

To generate responses from ChatGPT we used OpenAI's Python API using the chatGPT-35-turbo model. Despite it incurring a marginal cost, we specifically used the API in order to prevent the use of any text based data we collected to train OpenAI's core models. We provided to the model endpoint a prompt asking the model to generate a short reply (no more than 1 paragraph in length) to the text of one of the Reddit comments which had been replied to. We also provided the model with the content of the original post to ground the comment it was replying to within a larger discourse. We then collected the text it generated and converted it to word vectors using the same RoBERTa model we used for the Reddit comments. Our prompt is provided below in figure 1, and our OpenAI responses are available at https://tinyurl.com/23xda85x.

Not all comments could be passed to OpenAI due to tokenization window-length constraints. For this reason, we generated a total of 537 total ChatGPT generated replies.

```
Pretend that you are a community member
for the subreddit {subreddit}. You are
writing a reply to a comment on a post.
The original post has the following text:
"{original_post}". Respond to the following
comment: "{comment}" in the style of an
active member of this subreddit. Keep your
response to at most one paragraph long.
```

Figure 1: Example prompt used to generate ChatGPT responses.

Methods

Convergence-Entropy Measurement To measure the degree of convergence between either a human reply to a comment or an LLM generated reply we leveraged the semantic convergence-entropy metric (also called the Entropy-conVergence Metric, or EVM) first described in (Rosen & Dale, 2023). In that paper, the authors codify a convergence measurement based on the probability that the semantic meaning of each word in a sentence is represented in the textual makeup of some other sentence that it is compared to. They then calculate an entropy value for the entirety of a sentence when running this comparison in an effort to quantify how much of the semantic content of one sentence you might be able to predict after having read the other sentence. Sentences that have lower entropy are said to have higher convergence, because more of the content from the first sentence could be predicted from reading the second sentence.

Importantly, they show that this process can be efficiently approximated using contextually informed word vectors. This solves a number of data sparsity issues that would stand as a road block to doing this kind of study via other tools. To do this, they first calculate the probability that the semantic meaning of a word in one sentence, as represented by its contextually informed word vector, is captured by the use of any of the words in a second sentence. This is done by using a half-Gaussian prior over cosine error (CoE) values. This process, which is formally shown in equation 1, effectively represents under what conditions we ought to accept the statement that two word vectors code for the "same idea."

$$P(E_{xi}|E_y) = P_{\mathcal{N}_{[0,\infty]}}\left(\min_j\left(CoE(E_{xi},E_y)\right)\middle|\mu = 0.,\sigma\right) \quad (1)$$

They then use the values returned in equation 1 to calculate the Shannon entropy for the entire sentence by summing the entropy for every word *i* in the sentence.

$$H(x;y) = -\sum_i P(E_{xi}|E_y)\log P(E_{xi}|E_y) \quad (2)$$

While the authors describe equation 2 as capturing how much of the semantic content of one sentence can be predicted upon reading a separate sentence, there is another way to look at this measurement in the current context. In his preface to Shannon's essay in *A Mathematical Theory of Communication*, Weaver describes the condition under which Shannon entropy is high as exhibiting "a large degree of randomness or of choice" (Shannon & Weaver, 1949). This observation holds true for any estimation of Shannon entropy. In other words, we might reasonably expect EVM to increase in replies to a specific comment when the author of the reply

exerts a greater number of degrees of freedom in the construction of their reply. Put another way, the more that the author of the reply decides *not* to follow along with the conceptualization laid out in the comment they are replying to, the higher the entropy captured in the EVM will be.

In our specific experiment we are only interested in the degree to which the original comment that either humans or ChatGPT responded to can be recovered from the semantic content of that reply. For reference, then, let x be the original comment as written by a human Redditor, while y refers to replies which were written by either humans or ChatGPT.

Testing differences between conditions We employed two separate analytical regimes in order to better understand the degree to which ChatGPT replicated human EVM.

We tested whether or not the differences between the model's generated replies and human replies were due to random noise (the null hypothesis). Upon review of the data however, it became clear that there was a high degree of skew in at least entropy arising from human replies to comments, and possible within the chatGPT synthesized replies. Thus, we test first the degree and significance of the skew for both distributions, and then the difference between the two distributions via the Kruskal-Wallis H-test to assess gross differences between the two distributions in lieu of a parametric test.

We then isolated the relative contribution of a number of factors to the total convergence-entropy score in order to ascertain what was the total contribution in bits of entropy contributed by the difference between the LLM and human participants via linear mixed effects modeling. By carefully isolating out other potential factors that could arise from the speaker and other social factors, the value for the contribution of the LLM to the total convergence-entropy can thus be read as the contribution of illocutionary intent – the primary difference in the linguistic performance of human participants and the LLM. Importantly, we group our values according to the $y_{comment}$ – the reply to a previous comment as generated by either a human (actual replies) or ChatGPT (simulated replies). Every $x_{comment}$ or preceding comment was written by an actual human Redditor. We specifically isolate out the contribution of the number of upvotes that the original comment x received, the identity of the author of the comment x, entropic differences that could arise purely from the difference between subreddits (i.e., we assume there is a base rate of entropy and that the base rate varies from group to group per the findings reported in Rosen & Dale, 2023), and finally the contribution arising purely from whether the reply was written by a human or ChatGPT.

Results

As indicated above, we find that there is high, positive skew in the human made replies to comments. The skewness for the human data was found to be 25.47. The skew however for ChatGPT generated results was much lower and approximately normal at .42. Given the high degree of skew for the human generated data, we decided to deploy a Kruskal-Wallis H-test in lieu of a T-Test to assess differences in average entropy. We find that there is a significant difference in convergence-entropy ($H = 892.34, p < 1e^{-9}$) with replies generated by the LLM in general having lower entropy than human replies. We plot the entropy for the LLM vs. human participants in the KDE plot displayed in figure 2.

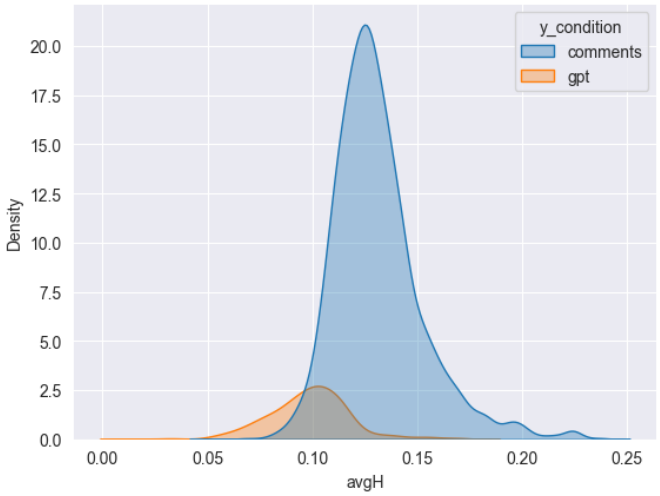

Figure 2: Kernel Density Plot for the average EVM value of each comment based on the reply for both human redditors (actual replies) and ChatGPT (simulated replies)

Additionally, we find a significant contribution of several factors to total convergence-entropy.

Unsurprisingly, the number of tokens in x is a significant contributor to total entropy. This is a given, even based on the original writings of Shannon (Shannon & Weaver, 1949). Greater numbers of tokens/events being measured will generally yield higher total entropy for the collection.

More interestingly, we find a significant contribution to total entropy arising from each subreddit community. Replies to comments in r/StarWars are characterized by 1.76 bits of additional entropy per comparison when compared to r/Dogs (r/Dogs being our de facto baseline). Replies to comments in r/StarTrek are characterized by an additional 3.00 bits of additional entropy per comparison when compared to r/Dogs. Humorously, this would point to conversations being potentially more contentious in r/StarTrek when compared to the other groups.

There is a significant contribution to total convergence-entropy arising from speaker identity for the author of the comment x ($2.18e^{-4}$). While that contribution is exceptionally small, it does indicate that there are individual differences with respect to who other users converge more readily too in online communities. Some folks just get converged to more readily than others.

No combination of factors with the number of likes received by the comment x contributed significantly to the total predicted entropy.

Var	coefs	stat	p
Intercept	-6.54	-10.01	1.31e-23
C(subreddit)[/StarTrek]	2.37	6.52	7.02e-11
C(subreddit)[/StarWars]	1.87	3.13	1.73e-03
x_comment upvotes	0.00	0.35	7.23e-01
is_LLM	-2.71	-9.77	1.48e-22
x_comment_upvotes x is_LLM	0.00	0.10	9.17e-01
n	0.14	412.91	1.48e-22
1‖x_user	0.03	12.98	1.57e-38
1‖x	0.01	2.42	1.54e-02

Table 1: LME results isolating the effects of all other user related variables and LLM/lack of illocutionary intent condition (`is_llm`).

Finally, we find that there is a significant contribution to entropy arising solely from whether a reply to a comment was written by a person versus if it was written by the LLM (-2.71). It is this value, again, that we attempted to isolate as a measurement of the influence of illocutionary intent, and its significance indicates that its contribution is not merely due to chance.

Our results are reported in Table 1.

After having isolated many possible sources of entropy arising from aspects of the identity of the various speakers, we interpret this main result in the following playful way: Relative to a generative language model lacking intentionality, human illocutionary intent adds, on average, 2.71 bits of entropy to the semantic content of a reply made to a comment. Compared to an average comment, where the median number of tokens in the comment x is 63 tokens, the median number of upvotes that an average comment is 73 upvotes, and the average contribution of the x user's individual convergence rates is 0.01, the illocutionary intent of the reply's author accounts for about 37.1% of the total entropy when trying to recover the ideas of the original comment from what is expressed in a reply to it in our data.

Illocutionary force and online engagement

Results in hand, let's discuss what they imply with respect to what human beings are doing when they interact with one another online. Most certainly, it is different from how our LLM "redditor" engages with content based on the entropic differences in the semantic meaning of utterances.

Why is it, thus, that people have such high entropy when conversing with one another? What is it about human interaction that lends itself to so much (apparent) noise? A few possibilities exist that might explain the difference observed. First, the difference could stem directly from the way that LLMs are trained. As we have already mentioned, LLMs are trained on large amounts of human text and generalize the transitional probabilities between words based on that training data. As pointed out by one of the reviewers, this could yield a situation in which an LLM learns to produce text that is in the "sweet spot" of informativeness – the outputs produced are predictable from the prior text, but do not outright mimic or parrot it. This could contrast with actual speakers, who may opt to construct utterances that maximize informativeness across contexts (?, ?), which would necessarily result in higher entropy.

As pointed out by the same reviewer, another reason could stem from the behavior of redditors themselves. Reddit users have a number of options for how to engage with a text beyond simply replying to it. They can also either upvote the text to indicate agreement, downvote the text to indicate disagreement, or simply take no action at all if they do not feel that they have anything meaningful to contribute (?, ?). Thus, the "noisiness" observed in human comments may just be due to only a subset of individuals replying to comments in the first place – individuals who both feel they have something meaningful to add.

Given that the objective of our study was to measure the degree to which *human* illocutionary intent may account for the semantic differences observed in online discourse, both of these points validate the use of LLMs as a foil to human commenters. Because of the way that they continue conversation, and because they seem to lack illocutionary intent, LLMs can act as a baseline for what one would expect a plausible comment to look like if we removed the human motivations for responding. If our objective is to generate an inhuman response, an unrestrained LLM is an appropriate tool for accomplishing that task.

So what is the *effect* of illocutionary force on how individuals engage with one another online? Here the difference between the entropy for the LLM when compared to human participants hints at an answer. We argued above that the LLM itself lacks illocution – an LLM, trained to complete a sequence of strings in a plausible way (?, ?), even with feedback about the acceptability of generated sequences (Brown et al., 2020), need not do so in a way that requires agentive intent. In fact, given that it generates tokens probabilistically, it is easiest to maximize the probability for the next token in a sequence by simply continuing the idea that was expressed in the previous turn.

Consider too the prompt used. While we prompt the model to create a string that plausibly sounds like another member of this subreddit community (a reasonable prompt given the extent to which contemporary language models are trained on Reddit data), the model ought to adopt the linguistic patterns of members of that community without needing to repeat the ideas expressed in the original comment a priori. But how does one add illocutionary force to a prompt? Or allow for variability in it? That is indeed an engineering question, and would be quite the feat to solve, given the current state of work in NLP with respect to both sociolinguistics and speech act theory. Indeed, perhaps advances in formal definitions for communication, like rational speech act theory (RSA; Frank & Goodman, 2012), may offer important clues. In any case, human communicators seem to infuse their messages with illocutionary intent. Even in a message board, there are a num-

ber of competing interests that shape a message based on how those interests speak to the speaker (?, ?). The difference between person and machine, at least at the time of writing this paper, could be one of illocution.

So what exactly does it mean that the illocutionary intent contributes 2.71 bits of entropy to redditors' comments? As one reviewer pointed out, there myriad ways that this entropic difference, even if it does hearken back to a difference in illocutionary intent, might manifest. Is it that the 2.71 bits is accounted for by more explicit reference to common ground? Could it be that human redditors are simply adverse to any form of redundancy? While these are interesting questions, to answer them requires a deeper dive into the qualitative content of responses – a task that, while important, we were not equipped to do for the current study (though we intend to follow up on it in subsequent work).

Conclusion

In the course of this paper we have done the following: (1) we established the importance of the illocutionary force in human communication, before (2) discussing how, based on their training objective, contemporary LLMs naturally seem to lack illocutionary intent. We then (3) used these two facts to set up an experimental design that allowed us to measure how much illocutionary intent contributes to the progression of conversations online. We did this via measuring the recoverability of the semantic content of comments from replies generated by both actual human participants and LLMs.

Admittedly, the contention that we are measuring the illocutionary force may seem somewhat comical. It is unlikely that J.L. Austin ever intended for his theoretical contribution to be situated within a quantitative framework. Even so, our results have important implications for the study of human communication. It is clear, at least in our experimental context, that when people converse with each other, they do much more than parrot each other's talking points. Indeed, controlling for illocutionary intent in conversation via the use of an intentionless LLM, what we see is that humans will systematically diverge from the ideas and semantic concepts of previous talking points in conversation when contributing their own utterances to ongoing discourse. That difference can be measured in convergence-entropy. By using an LLM as a proxy for the otherwise impossible condition of a interlocutor replying without illocutionary intent, we can, by comparison, measure just how influential illocutionary force is on conversational dynamics. The answer to that question is, humorously, that the illocutionary force exerts about 2.71 bits of entropy's worth of change in individual's responses.

More poignantly, demarcating the relative influence of an individual human's intentions on the content they generate during discourse could be a powerful counterpoint to some claims that humans are little more than simple statistical processes of this kind (and perhaps more so their *communicative practices*). In a 2020 paper, Bender and Koller proposed what they call the "octopus test" for language understanding. In it they describe a situation in which an interloper intercepts a message, and, by using the statistical regularities in that message, successfully returns to the sender something that appears to approximate an appropriate response (Bender & Koller, 2020). However, at no point is it necessary for the interloper to understand the original sender's intended meaning. Ultimately, a coherent message does not necessitate deeper understanding. We might mirror another precondition they allude to for the generation of agentive responses in human interlocutors: that the person authoring a response have some intent to do something at all with their speech act as opposed to simply responding for the sake of responding. This is consistent with several perspectives across decades of debate about human language, and is no doubt an exciting subject for continuing debate.

In their original paper, Rosen and Dale (2023) posit that entropic variation must be the rule as opposed to the exception (hence their "noisy Bernoulli principle"). But the source of that variance is left to experimenters to ascertain. They explored the effects of temporal distance, and social indicators of acceptance. We add to that another source of potential variance – illocutionary intent. Still, other social factors must also exist that affect the entropy between comments. And more research ought to be done to identify and catalog such factors.

References

Aurnhammer, C., & Frank, S. L. (2019, November). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107197–. Retrieved 2021-10-21, from https://www.sciencedirect.com/science/article/pii/S002839 doi: https://doi.org/10.1016/j.neuropsychologia.2019.107198

Austin, J. (1975). *How to Do Things with Words* (2nd ed.; J. Urmson & M. SbisÃ, Eds.). Harvard University Press.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. (arXiv: 2005.14165)

Byun, C., Vasicek, P., & Seppi, K. (2023). Dispensing with Humans in Human-Computer Interaction Research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–26). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3544549.3582749

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). *Does ChatGPT resemble humans*

in language use? arXiv. (arXiv:2303.08014 [cs]) doi: 10.48550/arXiv.2303.08014

Christiansen, M. H., & Chater, N. (2022). The unbearable lightness of meaning. In *The language game: how improvisation created language and changed the world.* New York: Basic Books.

Clark, R. (1974, May). Performing without competence. *Journal of Child Language*, *1*(1), 1–10. Retrieved 2024-05-10, from https://www.cambridge.org/core/journals/journal-of-child-language/article/performing-without-competence/BDD9 doi: 10.1017/S0305000900000040

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. (arXiv: 1810.04805)

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. doi: 10.1016/j.tics.2023.04.008

Enfield, N. J., & Sidnell, J. (2022). Intersubjectivity, Activity, Accountability. In *Consequences of Language: From Primary to Enhanced Intersubjectivity.* The MIT Press. doi: 10.7551/mitpress/14795.001.0001

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frankle, J., & Carbin, M. (2019). *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks.* arXiv. (arXiv:1803.03635 [cs]) doi: 10.48550/arXiv.1803.03635

Futrell, R., & Hahn, M. (2022). Information Theory as a Bridge Between Language Function and Language Form. *Frontiers in Communication*, *7*.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. doi: 10.1016/0167-2789(90)90087-6

Heintz, C., & Scott-Phillips, T. (2023). Expression unleashed: The evolutionary and cognitive foundations of human communication. *Behavioral and Brain Sciences*, *46*, e1. doi: 10.1017/S0140525X22000012

Holtgraves, T., & Ashley, A. (2001). Comprehending illocutionary force. *Memory & Cognition*, *29*(1), 83–90. doi: 10.3758/BF03195743

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis*. Psychology Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. (arXiv: 1301.3781)

Oller, D. K., & Griebel, U. (2021). Functionally Flexible Signaling and the Origin of Language. *Frontiers in Psychology*, *11*, 626138. doi: 10.3389/fpsyg.2020.626138

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/D14-1162

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. , 24.

Rosen, Z. P., & Dale, R. (2023). BERTs of a feather: Studying inter- and intra-group communication via information theory and language models. *Behavior Research Methods*. doi: 10.3758/s13428-023-02267-2

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press. (Google-Books-ID: IZ77BwAAQBAJ)

Snow, C. E., & Goldfield, B. A. (1983, October). Turn the page please: situation-specific language acquisition. *Journal of Child Language*, *10*(3), 551–569. Retrieved 2024-05-10, from https://www.cambridge.org/core/journals/journal-of-child- doi: 10.1017/S0305000900005365

Sperber, D., & Wilson, D. (2001). *Relevance: communication and cognition* (2nd ed ed.). Oxford ; Cambridge, MA: Blackwell Publishers.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008–3021.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models Know What Humans Know? *Cognitive Science*, *47*(7), e13309. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.13309) doi: 10.1111/cogs.13309

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. (arXiv: 1910.03771)

Zarcone, A., van Schijndel, M., Vogels, J., & Demberg, V. (2016, June). Salience and Attention in Surprisal-Based Accounts of Language Processing. *Frontiers in Psychology*, *7*. Retrieved 2024-04-30, from https://www.frontiersin.org/journals/psychology/articles/ (Publisher: Frontiers) doi: 10.3389/fpsyg.2016.00844