Using a state-of-the-art language model to probe mechanisms of language production:

Insights from GPT-2

An Honors Thesis for the Department of Psychology

Santiago Noriega

Tufts University 2021

Author Note:

Any correspondences should be addressed to the author's email address: airmak.g@gmail.com

Abstract

In the study of language production, Cloze and trigrams are among the most established methods to quantify the predictability of speech. Both of these methods have been extensively used to study language production in healthy subjects and people with neurocognitive disorders. However, both of these methods have limitations that impair their potential to be used as objective measures of psychosis. In this study, we introduce the use of a state-of-the-art machine-learning language model, GPT-2, as an alternative measure of speech predictability. By comparing GPT-2's probability calculations to Cloze and trigram probabilities for the same sample data, we found that GPT-2 can highly correlate to both measures. We then focused on the potential application of GPT-2 probability calculations to study speech production in schizophrenia, suggesting ways in which it can be used in place of n-grams and Cloze, and exemplifying how it can be used to test hypotheses on the underlying working memory deficits that have been linked to schizophrenia.

Keywords: GPT-2, Cloze, schizophrenia, language production, predictability

Introduction

Language offers humans a uniquely expressive method of communication. Successful communication through language requires both production and comprehension: one person produces language to convey a message and another person interprets the language produced in order to understand the message. Within this dynamic, the speaker must produce an output *coherent* enough for it to be reliably understood by the comprehender. Healthy adults produce highly coherent speech seemingly effortlessly, resulting in language output that is easily understood by others and robust to noise. At the same time, producing coherent language requires precise coordination of a number of high- and low-level cognitive faculties. Thus, studying language, and language impairments, can provide insights into the structure of the mind.

It is common for neuropsychiatric disorders to be coupled with abnormalities in the production or language. For example, schizophrenia is a neurocognitive condition where the breakdown of coherent language production is a common symptom. The speech of people with schizophrenia is characterized by its disorganized structure and often illogical flow of ideas, and at their most severe, these communicative abnormalities are known as *thought disorder* (Andreasen, 1979). One of the main clinical characterizations of thought disorder is *derailment*- where speech shifts between topics such that the preceding context may be unrelated to what comes next in the utterance. *Tangentiality* is another common linguistic symptom in schizophrenia, where a patient's speech is only tangentially related to the topic at hand. The following transcribed speech sample exemplifies how derailment and tangentiality may manifest in schizophrenia:

> "I always liked geography. My last teacher in that subject was Professor August A. He was a man with back eyes. I also like black eyes. There are also blue and grey eyes and other sorts, too…" (Bleuler, 1911/1950).

As it is evident in this speech sample, the syntax used by people with schizophrenia is normal. They are able to use and pronounce words as healthy people would, but the flow of the logic is incoherent.

Tangentiality and derailment, as markers of thought disorder, are often used as diagnostic measures for schizophrenia. However, the particular neurocognitive abnormalities underlying thought disorder are poorly understood (see Kuperberg, 2010 for review). As such, thought disorder diagnosis remains highly subjective and dependent on human recognition of these patterns, so a more objective paradigm to characterize the speech of schizophrenia patients has the potential to increase the scientific understanding of the cognitive deficits underlying the condition and improve its standard of diagnosis.

One way in which speech production *has* been quantitatively assessed in schizophrenia is through *predictability*. Note that here and throughout this study the term *predictability* is meant to specifically to describe *lexical* predictability. More coherent language tends to be more predictable, and conversely, incoherent language is usually very unpredictable. Thus, language production in healthy subjects tends to be highly predictable, since the speaker tends to make the deliberate effort to make themselves easily understood. Within the framework of predictability, words in an utterance can be assigned a probability value that indicates the likelihood that they appear following their corresponding context. For example, in the sentence "John mowed the lawn," the word *lawn* would likely be assigned a high predictability value since it is a very

probable continuation to the words preceding it (i.e. P(John mowed the lawn | John mowed the)).

Probability measures for predictability have traditionally been derived using the Cloze procedure (Taylor, 1953), which involves asking participants to perform a text continuation task based on what they think the next word is, given the context (see *Methods: Human Cloze Estimates* for description of the procedure). Their responses are then used to calculated probability distributions over possible continuations. The Cloze procedure remains the gold standard to quantify the predictability of words in a naturalistic context (natural discourse conversations) because it is a direct probabilistic measure of which words people consider to be the most probable continuations of a fragment (Felice and Buttery, 2019).

Whereas healthy people tend to produce highly predictable language, the manifestation of thought disorder as disorganized speech, as explained above, makes the linguistic output of schizophrenia patients more incoherent and, thus, measurably less predictable than that of healthy controls (Salzinger, 1964). To further understand this breakdown in language production, the Cloze procedure has been refined and modified for better use as a measure of communication problems in neurocognitive disorders like schizophrenia (Newby, 1998). Among these refinements, the responses given by participants are differentiated beyond the right/wrong binary to parse whether the answers were semantically similar to the actual continuation, or alternative but logical continuations. Cloze studies following these protocols have shown that the linguistic output of schizophrenia patients is measurably less predictable than that of healthy controls (Salzinger, 1964; Newby, 1998). Considering this evidence, proponents of the Cloze technique have considered Cloze to have the potential to be an objective psychopathological measure (Newby, 1998). However, these early studies on the predictability of speech by people with

schizophrenia had significant flaws. Cloze is time- and labor-intensive to collect, so these studies only collected Cloze probabilities from select words within passages, and they had relatively small sample sizes (e.g. 13 participants with schizophrenia in the Salzinger, 1964 study). Also, for such procedure to be realistically useful in diagnosis, it would need to be further automated so that numerous participants do not have to perform the Cloze protocol to analyze the predictability patterns of every potential patient.

Because Cloze is time- and labor-intensive to collect, it is less suitable for examining subtle differences in patients versus controls. An alternative approach, that may be used to approximate Cloze probabilities, is to use computational models to generate probability distributions from the speech of participants. Computational models also have the benefit of being automated, so they can produce data on speech predictability without the need for other participants to evaluate their speech (e.g. via the Cloze procedure). So, if a language model were to reliably reproduce Cloze probabilities, it could fulfill the promising capacity of an objective psychopathological measure without the impracticality associated with the Cloze procedure. This means that language models can offer a different perspective through which to understand language output in schizophrenia.

On the other hand, perhaps the simplest type of language model which can be used to estimate speech predictability is the n-gram model, which involves calculating the probability of a word completely based on the words that preceded it in the library used to train the model (Niesler and Woodland, 1996). The n-gram procedure involves breaking up a corpus of text, used to train the model, into an array of continuous word-by-word sequences grouped into *n* words. For example, given the sentence *"The dog was chasing the cat,"* a *tri*gram model would form the following trigrams from the sentence: (The dog was), (dog was chasing), (was chasing

the), (chasing the cat). This way, words become associated with other words that precede or follow them in the corpus, and the probability of their occurrence in a new context is calculated from their abundance in a context with the same words in the training corpus.

N-grams, particularly trigrams, have been extensively used to study probabilistic features of naturalistic speech in healthy adults. For example, Xu and Reitter (2018) used trigram models to test predictions of the uniformity of information density using the British National Corpus (BNC) as its library of naturalistic text. Numerous other researchers have also used trigram models to obtain predictability measures in healthy adults (e.g. Genzel & Charniak, 2001; Qian and Jaeger, 2011).

However, the accuracy of the n-gram procedure is extremely dependent on the richness of the library it is built on and does not consider the influence of context apart from the n-1 words preceding the target word. If out of chance, a particular critical word is never associated with a context word that nonetheless is highly predictable of the critical word, the n-gram model will not be able to distinguish this relationship. Relatedly, n-grams are context-fixed: that is, they make their probabilistic calculations based on a fixed amount of context that cannot be altered or used to make calculations outside the specific range calculated. This can be particularly problematic when studying questions for which different levels of context could be a relevant factor.

The purpose of this study was to investigate the ability of a machine-learning language model, OpenAI's GPT-2 (Radford et al., 2019), to calculate word probabilities as measures that resemble those produced by the better-established Cloze and n-gram methods. In other words, given a sentence fragment, GPT-2 can calculate the probability of each word's occurrence given the context of preceding words, and whether or not those probabilities resemble other probability

measures remained an open question. GPT-2, a transformers-based machine learning Language Model, is an unsupervised multitask learner, meaning that it does not need to be trained for task-specific purposes to perform multiple tasks related to language comprehension. Also, unlike n-grams or Cloze probabilities, GPT-2 can be given varying amounts of context to make predictions. As I will discuss, this makes it a potentially useful tool for studying language production in schizophrenia patients, but in order to use it for this purpose, the way variations in context affect GPT-2 probabilities calculated for the speech of healthy adults needed to be established.

Although the resemblance of word probabilities produced by GPT-2 to those generated by Cloze and n-grams remained open to investigation, its predictive power as a Language Model had been strongly established by different measures. For example, GPT-2's performance was remarkably close to that of human subjects in the Children's Book Test (Hill et al., 2015), a simple Question Answering Task and other tests of language comprehension and production (Radford et al., 2019). GPT-2 is therefore likely to greatly outperform n-gram models when measuring predictability, and it is likely to mimic the probabilities generated during Cloze procedures.

For these reasons, GPT-2 may prove to be a powerful tool for understanding how language production breaks down in schizophrenia. Thought disorder in schizophrenia has been extensively studied, but the underlying mechanisms governing these deficits in language production remain uncertain (i.e. Bleuler, 1911/1950). It has been suggested that working memory, the ability to hold information temporarily for executive recall, is defective in schizophrenia patients, causing problems for them as they try to build up and use context to form new utterances (reviewed at length in Kuperberg, 2010). The evidence supporting this theory

includes clear indications of working memory deficits in schizophrenia patients, which leads to an overall poor performance in executive function tasks (Lee and Park, 2005). For example, severity of thought disorder has been positively associated with performance on the Stroop task, which tests executive function (Kerns, 2007). This association has been corroborated in a meta-analysis by Kerns and Berenbaum (2002). Still, neuropsychological batteries that include tasks like the Stroop task have the disadvantage of possibly conflating multiple cognitive processes into single performance measures. Furthermore, it has also been shown that giving participants more context of a speech sample from a person with schizophrenia during the Cloze procedure did not help them predict the speech of people with schizophrenia better, suggesting that schizophrenia patients may have trouble using the context of their previous utterances to produce comprehensive language (Salzinger, 1970).

By using GPT-2 to calculate the probabilities of words uttered by schizophrenia patients in a controlled interview setting, we can determine if they appear to indeed lack the ability to effectively build upon preceding context to make upcoming words as predictable as they would be with an intact working memory. To test if poor working memory underlies thought disorder in schizophrenia, GPT-2 can be used to quantify whether increasing context increases the predictability of upcoming words in the speech of people with schizophrenia as much as it does in the speech of healthy subjects. As previously discussed, this can be tested by using GPT-2 because of its flexibility to consider varying amounts of context preceding a critical word, meaning that distinctions can be made on the predictability of the speech of people with schizophrenia when only a local context or a more global context was used to utter a word.

The first step to be able to use GPT-2 for the purpose of studying speech in people with schizophrenia is to characterize how well the model estimates the predictability of speech in

healthy adults in naturalistic conversations. To do this, we conducted two experiments. First, we used GPT-2 to obtain predictability measures from a corpus composed of four-sentence vignettes designed to resemble naturalistic sentences (refer to *Methods- Corpus 1: Semi-naturalistic Vignettes*), and analyzed how they resembled analogous Cloze and n-gram derived measures. Second, we calculated the same predictability measures with GPT-2 across varying levels of context to test how well the context-dependent patterns of these measures line up with what would be expected of a model that improves its measures as it is given more context. These are important steps to establish a baseline of how GPT-2 interprets naturalistic speech and calculates relevant predictability measures, and they will provide a starting point for comparison when the language model is used to interpret naturalistic speech by people with schizophrenia.

We aim to provide a protocol and proof of concept for the use of GPT-2 as a practical alternative to well-established measures of lexical predictability. As discussed above, the setbacks of n-gram and Cloze procedures limit their potential use as objective psychopathological measures. Here, we will focus on the potential application of GPT-2-derived measures of predictability in schizophrenia, suggesting ways in which it can be used in place of n-grams and Cloze, and exemplifying how it can be used to test hypotheses on the underlying working memory deficits that have been linked to schizophrenia.

Methods

GPT-2

GPT-2 (Radford et al., 2019) is a machine-learning language model developed by OpenAI. The novelty of GPT-2 as a predictor of human language comes from its Transformer architecture (Wolf et al., 2020) and the large library of text on which it was pre-trained. GPT-2 was trained on WebText (Gokaslan and Cohen, 2019), a diverse library of over 8 million human-curated online documents of 40GB in total.

In this study, we used the pre-trained GPT2-Large model from HuggingFace's Transformers repository (Wolf et al., 2020), which has 36 layers, 345 million parameters, and a vocabulary of 50,257 tokens. GPT2-Large was used to calculate probability distributions over possible continuations for a variety of contexts. These probability distributions were extracted using the *GPT-2-for-Psycholinguistic-Applications* repository (github.com/samern92/GPT-2-for-Psycholinguistic-Applications). For example, for the incomplete sentence "*John mowed the …"* GPT-2 formulates a probability distribution of an *n* number of possible continuations given "*John mowed the "* as the context fragment. If the actual continuation to the fragment in this case is *lawn*, we would extract the probability of *"lawn"* being the next word from the probability distribution generated by GPT-2. These distributions were used to compute measures analogous to those used in human-generated Cloze studies, as well as entropy as an information theoretic measure.

Probability distributions were calculated for varying levels of context, for each sample of text, by manipulating the number of words before the target word which were given as the context to GPT-2. Using the same sample fragment "*John mowed the …"* followed by the word

lawn, we could determine how different levels of context fed to GPT-2 affect the probability of "*lawn*" being the continuation of the fragment. For example, GPT-2 could be manipulated to find the probability of *"lawn"* given the two preceding words, P(*mowed the lawn | mowed the*), the whole preceding sentence, P(*John mowed the lawn| John mowed the*), or any other level of context for up to 1024 preceding tokens.

Predictability Measures (GPT-2, n-gram, and Cloze probabilities)

GPT-2's performance as a tool for calculating predictability in naturalistic language was compared to the longstanding Cloze and n-gram approaches to calculating word probabilities. The n-gram modeling technique involves dividing a text into n-word components, building a Markov chain of order where the probability of a word depends on the words that preceded it in the input library of text (Biemann et al., 2015). In the case of this, *tri*grams would split a library of text into three-word components that would be used to approximate the probability of occurrence of a word given the two words that came before it. Trigram language models were formed using the Python Natural Language Toolkit (NLTK) package (Bird et al., 2009), and Kneser-Ney smoothing was applied to the models (James, 2000). The Cloze procedure (Taylor, 1953) involves calculating probabilities from the answers of participants who are asked to perform a text continuation task based on what they think the next word is (refer to *Data Collection* for protocol).

From GPT-2 probabilities, measures of constraint, entropy, and surprisal were derived. Constraint is the probability of the most likely continuation in the probability distribution calculated by GPT-2 when given a preceding context. Entropy if a measure often used as a proxy

for information density (Shannon, 1948). In a probability distribution *X*, as calculated by GPT-2 for *n* possible continuations given the context, the entropy of *X* is x_i:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where the probabilities of each word x_i were normalized within their probability distribution.

Human Cloze Estimates

Cloze measures for each word in *Corpus 1* (see description below) were crowdsourced using Amazon Mechanical Turk by presenting workers with one word of the sentence at a time and asking them to fill in the following word. For the sentence above, the first three set of stimuli presented to workers would be as follows, with the workers being instructed to continue the sentence with a one word continuation to fill in the blank.

1: The ____

2: The boy ____

3: The boy scouts ____

(Refer to the Cloze Sentence Completion Task described by (Schnoebelen and Kuperman, 2010) for a complete description of the methodology).

Stimuli were divided up into nine non-overlapping lists, each of which was posted as a separate task (HIT) on mTurk. Individual mTurk workers were allowed to complete more than one list, but were not allowed to complete any particular list more than once. Workers were eligible for the study if they had an IP address within the United States, had completed 100 or more HITs on Amazon Mechanical Turk, had an overall HIT acceptance rate of greater than

98%, and were native English speakers. This resulted in the collection of 477 datasets (mean per list = 52.56). Workers gave written informed consent prior to participating in the study, in accordance with the procedures of the Institutional Review Board of Tufts University.

Of these 477 datasets, 4 were excluded for being less than 50% complete. Another 96 datasets were excluded for poor accuracy (<10%). This threshold was determined by examining the distribution of accuracy scores across participants, which was strongly bimodal; manual examination of the datasets with less than 10% accuracy revealed that few of these responses appeared to be made in good faith (rather, low accuracy tended to result from participants entering nonsense responses, e.g. random strings of letters or the same word over and over). An additional 6 datasets were excluded for unusually high accuracy (accuracy > 50%), which suggested that participants had done the experiment more than once. Amongst the remaining 361 datasets (185 female, 168 male; mean age = 33.54, SD = 8.76) mean accuracy was 25.15% (SD = 4.31%).

Each (word-by-word) response was screened for spelling errors using the Python library PyEnchant; responses that were flagged as misspelled were checked manually. When it was clear what the misspelled word was supposed to be (e.g. “suprise” for “surprise”), the response was corrected. If it was unclear, the response was left as is. Responses which were not word-like (e.g. strings of random letters) were excluded from the data. After cleaning, each trial had an average of 40.06 individual responses (SD = 3.91, min = 28, max = 46). These data were used to calculate cloze probability, constraint, and entropy for each trial.

Calculation of response times for each trial required some further cleaning. The very first trial of each experiment was excluded, as were any trials where the first letter the participant typed did not match the first letter of their final response and any trials where the response time

(time for the participant to begin typing) was longer than ten seconds. Finally, we excluded response times which were outside two standard deviations of the mean for each trial. This resulted in an average of 36.01 (SD = 3.53, min = 18, max = 45) response time measurements per trial.

Corpus 1: Semi-naturalistic Vignettes

The first corpus we analyzed was made up of 179 vignettes of four sentences each. These stories were designed to resemble naturalistic sentences that could have been formed in an everyday conversation, with varying degrees of predictability. Each sentence had between eight to fourteen words, and all the sentences related to the others within the same set. For example:

"The boy scouts were excited about their hiking trip. They were going to spend an entire week in the wilderness honing their skills. They had been on similar adventures before and knew just what to do. After arriving in new territory, the scouts would always explore the area."

For each word in *Corpus 1*, GPT-2 probabilities and derived measures (constraint, entropy, and surprisal) were compared with methods obtained from the Cloze task and trigram language modeling by Pearson correlations between analogous measures. When GPT-2 output was compared with Cloze measures, GPT-2 was given the entire context within a vignette preceding a target word, since this would be the same amount of context seen and considered by participants during the Cloze task. When GPT-2 output was compared with trigram model measures, GPT-2 was given the two words preceding the target word as the context.

From the experimentally designed semi-naturalistic vignettes, each of which consisted of four sentences of similar length, Cloze probabilities were crowd-sourced using Amazon

Mechanical Turk responses (Alexander, in print). From these Cloze probabilities, entropy measures were derived as well.

From the same semi-naturalistic vignettes, GPT-2 Large was used to obtain probability measures for each of the words. The GPT-2 model was designed to either consider either the entire context within a vignette preceding a word (see *GPT2 Prob.* in Table 1), or to only consider the sentence a word is on as its preceding context (see *GPT2 Sentence Prob.* in Table 1). For example, given the fragment "*The boy scouts were excited about their hiking trip. They were going […],"* data was collected from the GPT-2 model to either consider the entire preceding fragment as the context for the word *"going"* (*GPT2 Prob.*), or to only consider *"They were"* as the context (*GPT2 Sentence Prob.*). From both of these levels of GPT-2 probability measures, entropy was calculated as well (*GPT2 Entropy* and *GPT2 Sentence Entropy* correspondingly, in Table 1).

Corpus 2: Natural Spoken Language

The British National Corpus (BNC, BNC, 2007) is an extensive and diverse collection of written and spoken language of British English. The spoken section of the BNC was compiled to resemble a representative sample of spoken British English and is further subdivided into the DEM-BNC and the CG-BNC (Burnard, 2007). The DEM-BNC is a "demographically-sampled" collection of everyday, informal conversations among volunteers of different demographics (Aston and Burnard, 1998). The CG-BNC is a "context-governed" collection consisting of more formal conversations or speeches ranging generally centered around a specific topic (Aston and Burnard, 1998). For example, a casual conversation between friends would be representative of

DEM-BNC transcript, whereas subjects covered in the CG-BNC range from soccer punditry to history lectures.

The DEM-BNC makes up about 40% of the Spoken corpus, and the CG-BNC the remaining 60% (Aston and Burnard, 1998). In total, thousands of speakers and 4.2 million words are transcribed in the Spoken BNC (Aston and Burnard, 1998). The BNC also includes XML markers corresponding to the speaker saying each word (Burnard, 2007). This allowed us to define an utterance as the words said by a single speaker without interruption and to use it as a level of context.

The BNC was divided into the CG-BNC and the DEM-BNC sub-corpora, and for each of them, GPT-2 was configured to consider either the preceding two words (*trigrams*), the preceding sentence, the preceding words spoken continuously by the same speaker (*utterances*), or the full context preceding a critical word, and GPT-2 probabilities were calculated for each word at each of these different levels of context. Thus, a library was accumulated where every word in the spoken section of the BNC was attached to GPT-2 probability measures across different levels of context, along with markers for sentence boundaries, changes in speakers, and changes in transcripts.

Perplexity was calculated from the data collected from the BNC corpora as a measure of how well GPT-2 is able to predict the identity of upcoming words in a text. Lower perplexity indicates probability distribution that is better at predicting words in a sample. Mathematically, where w_1, w_2, … w_N are all the words in a corpus, and $p(w)$ is the GPT-2 probability of a given word, perplexity P was calculated as:

$$P = b^{\left(-\frac{1}{N}\right)*\sum_{i=1}^{N}\log_b p(w_i)}$$

In order to perform the GPT-2 calculations necessary to obtain probabilities associated with every word in the BNC corpora, across multiple levels of context, the Tufts University High Performance Computing (HPC) cluster servers was used. Each BNC sub-corpus was split by the individual transcripts that comprise them, and the GPT-2 calculations needed were run in parallel computational nodes, which was necessary in order to obtain these data from corpora of this magnitude within a reasonable amount of time.

Corpus 3: Interview transcripts

Speech samples from standardized interviews between an experimenter and either a schizophrenia patient (n = 1) or healthy control subject were transcribed (n = 32). At the current stages of this study, only one transcript from the interviewed participants with schizophrenia has been transcribed, although the recordings of such interviews are available and are currently in the process of transcription. Thus, the main goal of using these transcripts was as an exploration of a procedure that will be followed-up with a greater sample of interviewed people with schizophrenia. These transcripts were divided into texts consisting of subject utterances containing (*A*) 20 words or more, (*B*) 15 words or more, or (*C*) 10 words or more (Fig 9). GPT-2 Large was used to extract probabilities from each word on each of these sub-transcripts, with the context level considered by GPT-2 being ranged between 1 and 15 words.

Results

Corpus 1: Semi-naturalistic Vignettes

To assess whether GPT-2 measures emulated the patterns in word probability distributions obtained from the crowd-sourced Cloze technique, GPT-2 probabilities derived from the full preceding context and from a sentence-level-only context before the critical word were compared with the corresponding Cloze measures for each word in the vignettes using Pearson correlations (Table 1). Highlighted below in Table 1, several of these correlation coefficients suggested that the GPT-2-derived calculations are accurate at measuring the same patterns in probability as the Cloze technique. First, there were strong correlations between Cloze probabilities and both the full-context and sentence-context GPT-2 probabilities ($r = 0.80$ and $r = 0.73$ respectively). This correlation was significantly higher when the GPT-2 model considered more words as part of the context, the same words the participants performing the Cloze task saw as context, further suggesting that GPT-2 can perform as an appropriate proxy measure for Cloze probabilities. As expected, GPT-2 probabilities were strongly inversely correlated with GPT-2 entropy derivations (see red highlights, Table 1), while being strongly correlated with Cloze entropy derivations, suggesting that entropy as a GPT-2 measure is also an accurate representation of Cloze-derived entropy.

Qualitatively, GPT-2 and Cloze probability-obtaining methods appear to be sensitive to the same patterns over the course of sentences and vignettes. As shown in Fig 1, rises and drops in word probabilities averaged by the word's position in a sentence are mirrored by GPT-2 and Cloze, particularly when the GPT-2 model considers the same amount of context that Cloze task participants did. This figure also suggests that GPT-2 probabilities, thought they follow the same

patterns, are generally lower than Cloze probabilities. Here, it is also observed that GPT-2 probabilities that only considered the local sentence as context were consistently lower than those obtained from the model that considered the full vignette context, which was expected. As other results will show, it was a reoccurring feature of GPT-2 that its calculations for probability increased as the amount of context it was given increased.

Entropy calculations from GPT-2 are similarly good predictors of Cloze-derived entropy. The vignettes were split up into each of their four sentences, and then grouped by their position in the vignette (first sentence, second sentence, third sentence, and fourth sentence), and entropy measures were averaged for each sentence grouping by a word's position in the sentence in order to track how GPT-2 entropy measures compare with the Cloze derived measures (Fig 2). When averaged across all sentences by word position, the trend similarities are not as evident, although a similar broadly decreasing trend in entropy as a sentence progresses is observed by both types of calculations (Fig 3).

It was then assessed whether trigram probability measures could also be emulated by GPT-2. In order for GPT-2 to most closely resemble the calculations performed by a trigrams language model, the amount of context that GPT-2 was given preceding a critical word was shortened to two words in the *GPT2 Trigrams* condition (Fig 4). The correlation between *GPT2 Trigrams* probability outputs and traditional trigram probably calculations was high (r = 0.810), and as in the GPT-2 comparison with Cloze probabilities, rise and drop patterns were remarkably similar in the two measures (Fig 2).

Table 1. Correlation constant *r* values between GPT-2 and Cloze measures. Cloze derived probability and entropy measures are labeled in gray and GPT-2 derived probability and entropy measures are labeled in shades of blue.

	Cloze Prob.	*Cloze Entropy*	*GPT2 Prob.*	*GPT2 Entropy*	*GPT2 Sentence Prob.*	*GPT2 Sentence Entropy*
Cloze Prob.	1					
Cloze Entropy	-0.67162	1				
GPT2 Prob.	0.802896	-0.59478	1			
GPT2 Entropy	-0.60697	0.715363	-0.73567	1		
GPT2 Sentence Prob.	0.729564	-0.56275	0.888749	-0.68058	1	
GPT2 Sentence Entropy	-0.56705	0.654371	-0.66846	0.855017	-0.75236	1

Fig 1. Word probabilities measured by GPT2 or the Cloze technique, averaged by their position in a sentence. X axis: position of a word in a sentence (i.e: the word "*dog"* would be in position 2 within the sentence "*The dog was chasing the cat"*). Y axis: average probability measured across all words per position in a sentence. Standard error bars shown.

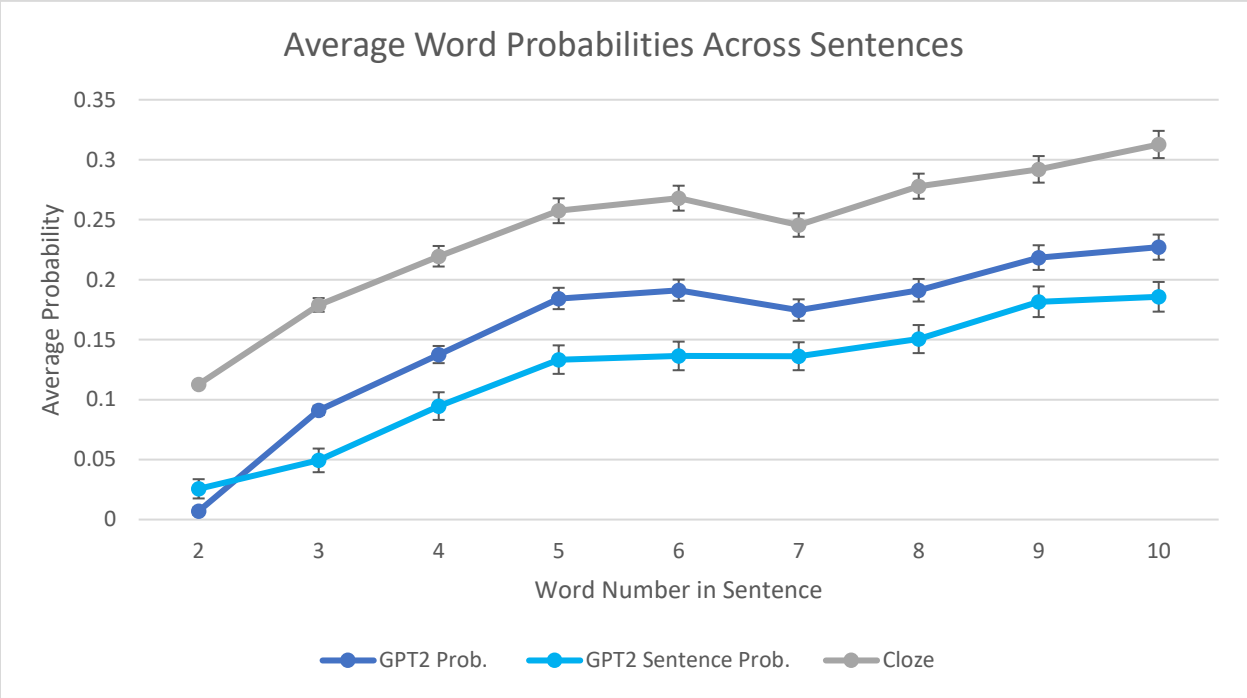


Fig 2. Clause entropy measures obtained by GPT2 (solid lines) or the Cloze technique (dashed lines). Each of the four sentence positions within a vignette are marked by a specific color. Dark blue: sentence 1; Red: sentence 2; green: sentence 3; yellow: sentence 4.

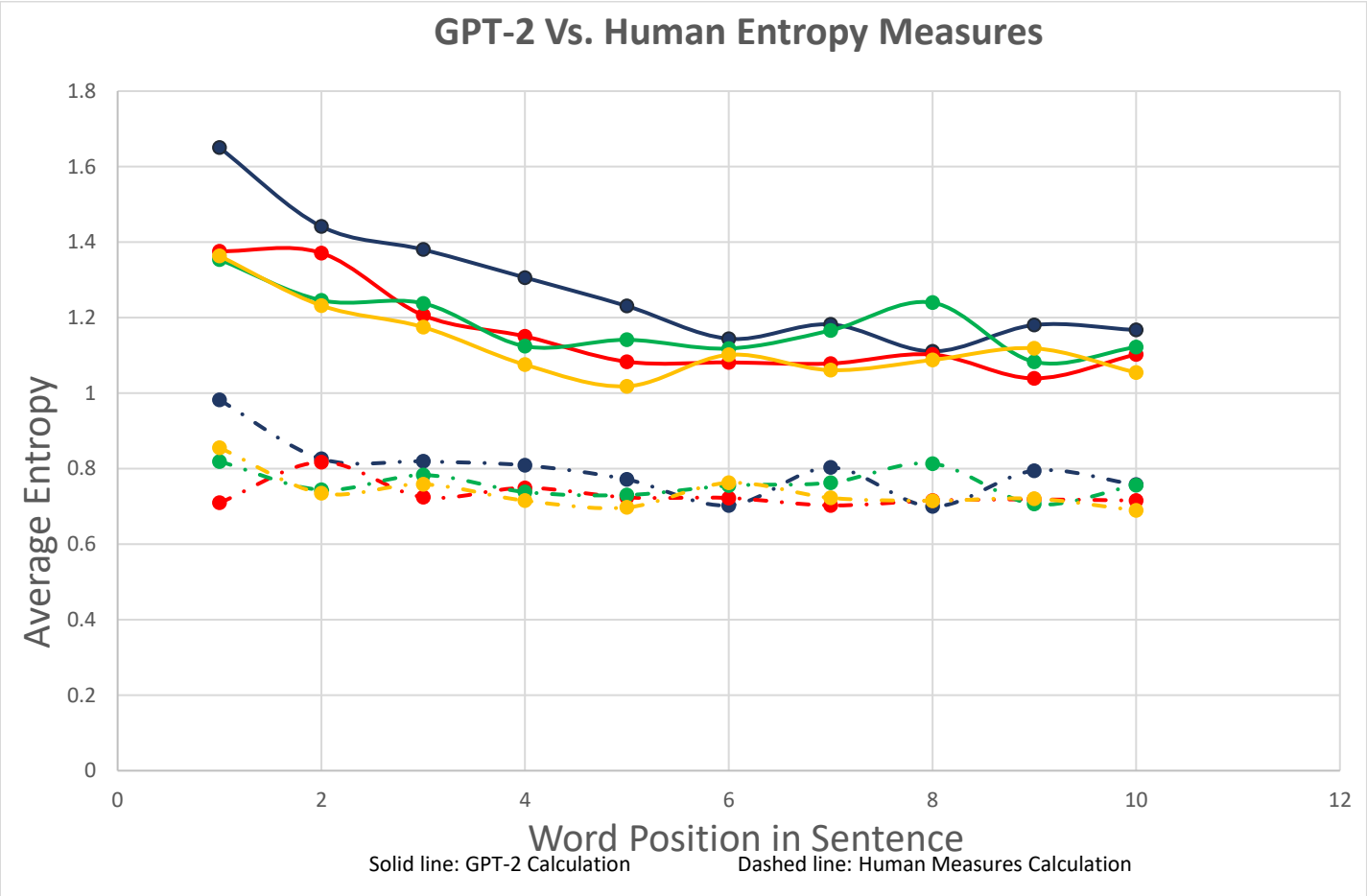


Fig 3. Clause entropy measures obtained by GPT2 or the Cloze technique, and averaged by their position in a sentence. Dashed lines represent the linear regressions of each curve. Standard error bars shown.

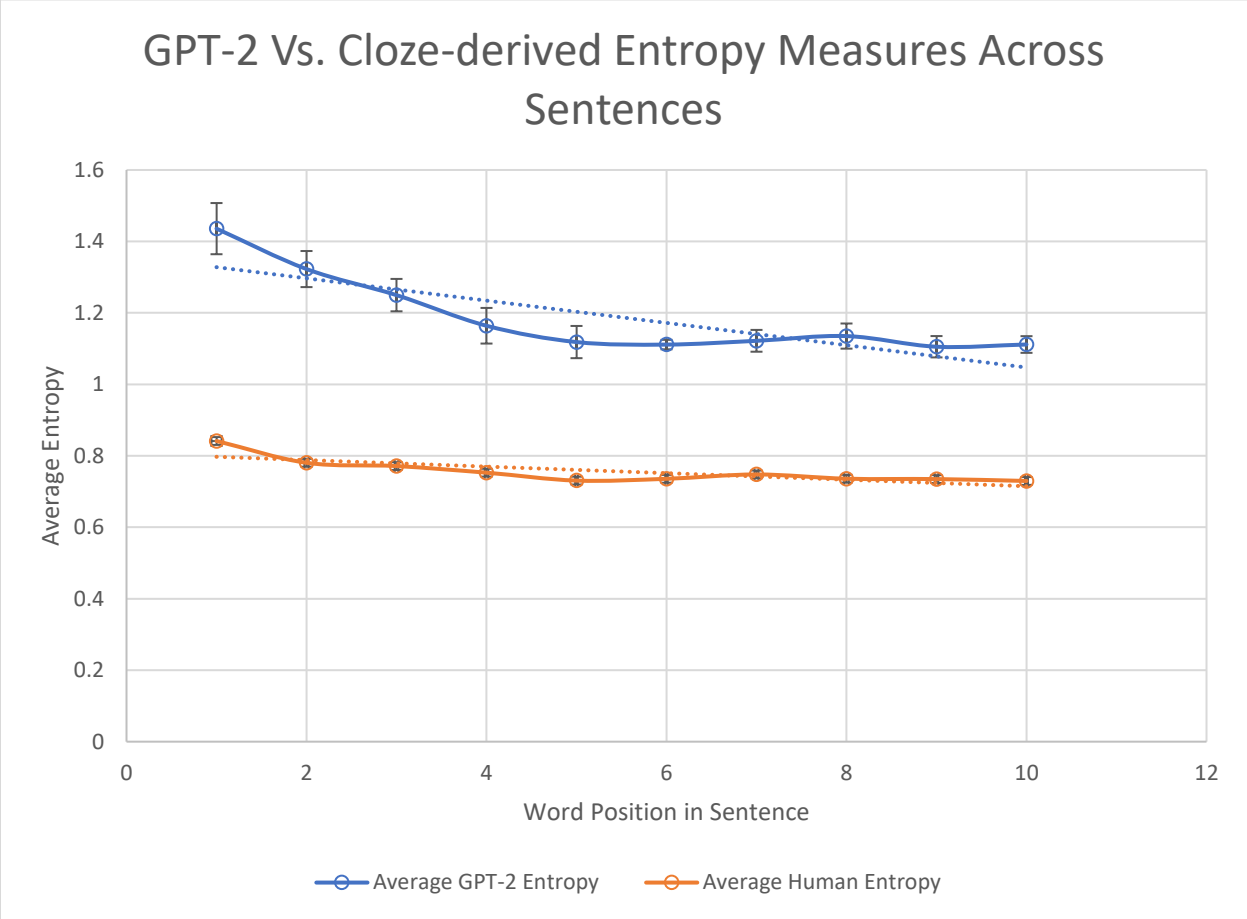
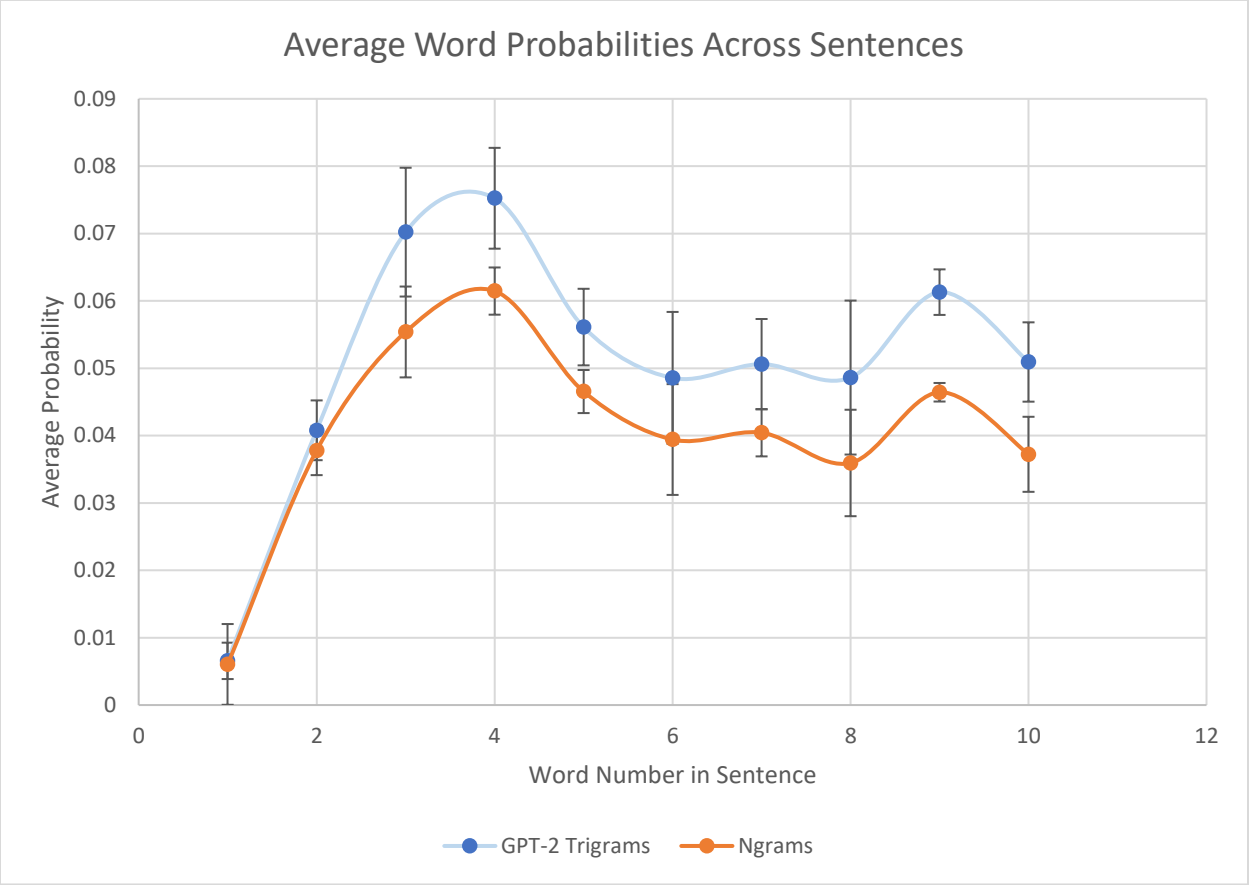


Fig 4. Word probabilities measured by GPT-2 or a trigram model, averaged by the word's position in a sentence. Standard error bars shown.



Corpus 2: Natural Spoken Language

Then, to assess GPT-2's performance as a measurer of predictability in naturalistic speech, GPT-2 Large was used to obtain probability measures for every word in the spoken section of the British National Corpus (BNC), across different levels of context. Because language models are traditionally compared using *perplexity* (see *Methods: Corpus 2* for description), rather than raw probabilities, we here report perplexity values along with the probabilities. If GPT-2 adapted well to "learning" from spoken words, it would be expected that as more context was evaluated by GPT-2, the perplexity calculated across a transcript would decrease. Indeed, for both the CG-BNC and the DEM-BNC, as context grew larger, average perplexity reliably decreased (Fig 5 and Fig 6). Interestingly, perplexity for the DEM-BNC was greater than perplexity for the CG-BNC at all levels of context. Since perplexities were calculated from the raw probabilities, the inverse pattern was observed as expected from both the CG-BNC and the DEM-BNC, with word probabilities averaged across each of the corpora increasing with increasing context consideration by the model (Fig 7 and Fig 8). There was once again a shift observed between the averaged measures in the DEM-BNC and the CG-BNC- mean probability was higher across all context levels in the CG-BNC, implying that GPT-2 was better at internally predicting upcoming words in the CG corpus.

Fig 5. Differences in perplexity of GPT-2 probability calculations by amount of context considered by GPT-2 in the CG-BNC corpus. Standard error bars shown.

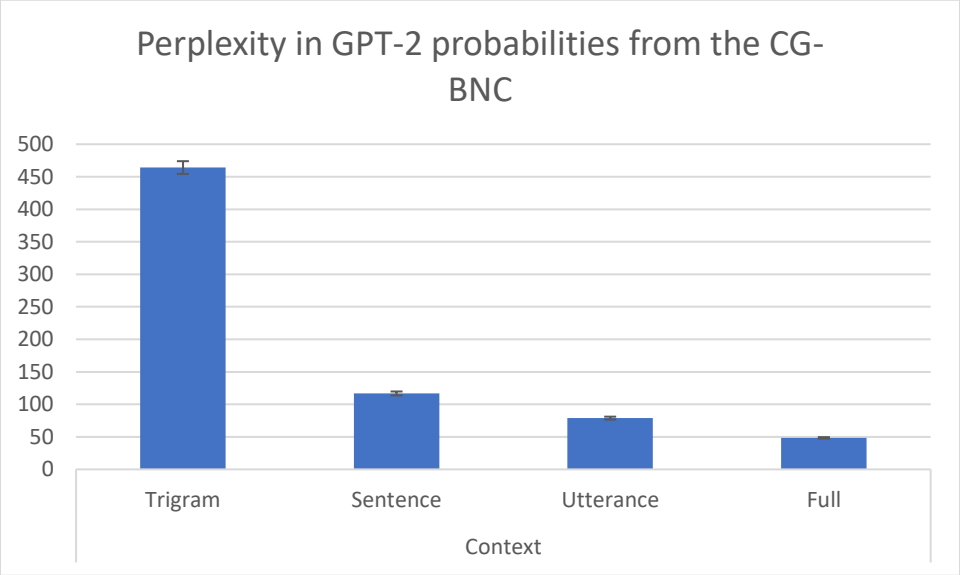


Fig 6. Differences in perplexity of GPT-2 probability calculations by amount of context considered by GPT-2 in the DEM-BNC corpus. Standard error bars shown.

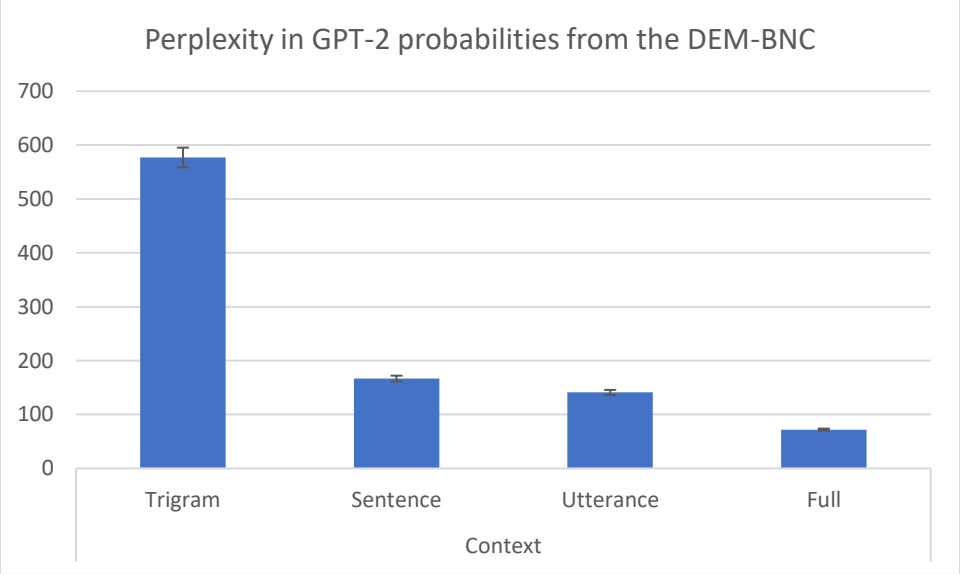


Fig 7. Differences in mean probability by amount of context considered by GPT-2 in the CG-BNC corpus. Standard error bars shown.

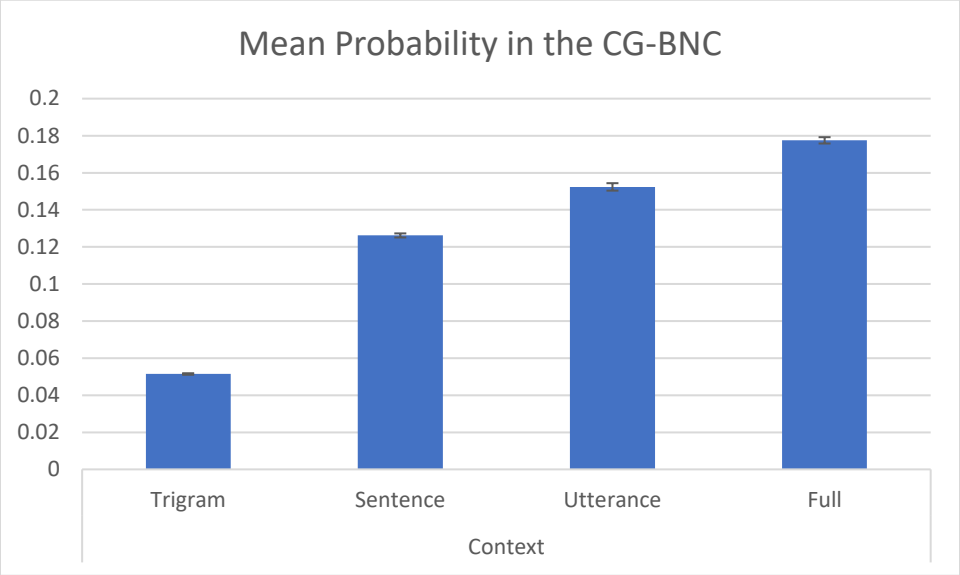
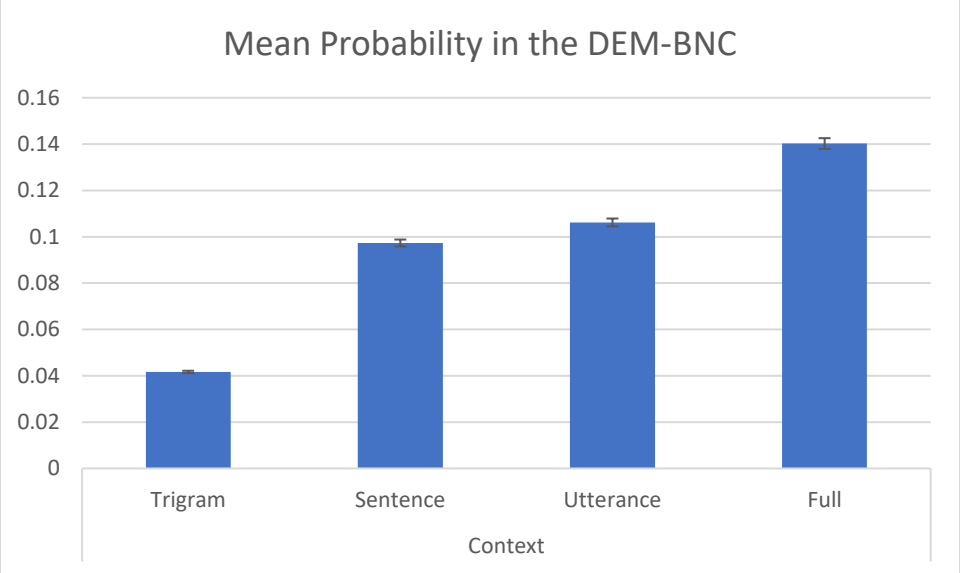


Fig 8. Differences in mean probability by amount of context considered by GPT-2 in the DEM-BNC corpus. Standard error bars shown.



Corpus 3: Interview Transcripts

With the accumulated evidence of GPT-2's ability to obtain reliable probability measures from naturalistic speech as indicators of word predictability, we sought proof of concept that GPT-2 can be applied to distinguish patterns in speech predictability between people with schizophrenia and controls. For all of the sub-transcripts of these interviews (refer to *Methods: Interview transcripts*), separated by a minimum threshold of words per utterance for the utterance to be included, GPT-2 probability increased as the amount of context considered by GPT-2 increased (Fig 9). This increase was nearly linear for all three sub-transcripts (r^2= 0.965, 0.961, and 0.950 for sub-transcripts *A*, *B*, and *C* respectively). Furthermore, there is a relatively small but consistent upward shift in GPT-2 probabilities for the sub-transcripts that included longer utterances.

The standard deviations of probabilities calculated by GPT-2 across all context levels for each of the sub-transcripts were obtained and compared (Fig 10). Consistently, as the number of words that were considered by GPT-2 as context increased, the standard deviation of GPT-2 probabilities within a sub-transcript also increased. Once again, there is a relatively small but consistent upward shift in standard deviations for the sub-transcripts that included longer utterances.

Fig 9. Average GPT-2 probabilities across transcripts by different amounts of context given to the model. The speech sample is from standardized interviews given to schizophrenia patients and health control. *A*: subject utterances 20 words or longer. *B*: subject utterances 15 words or longer. *C*: subject utterances 10 words or longer.

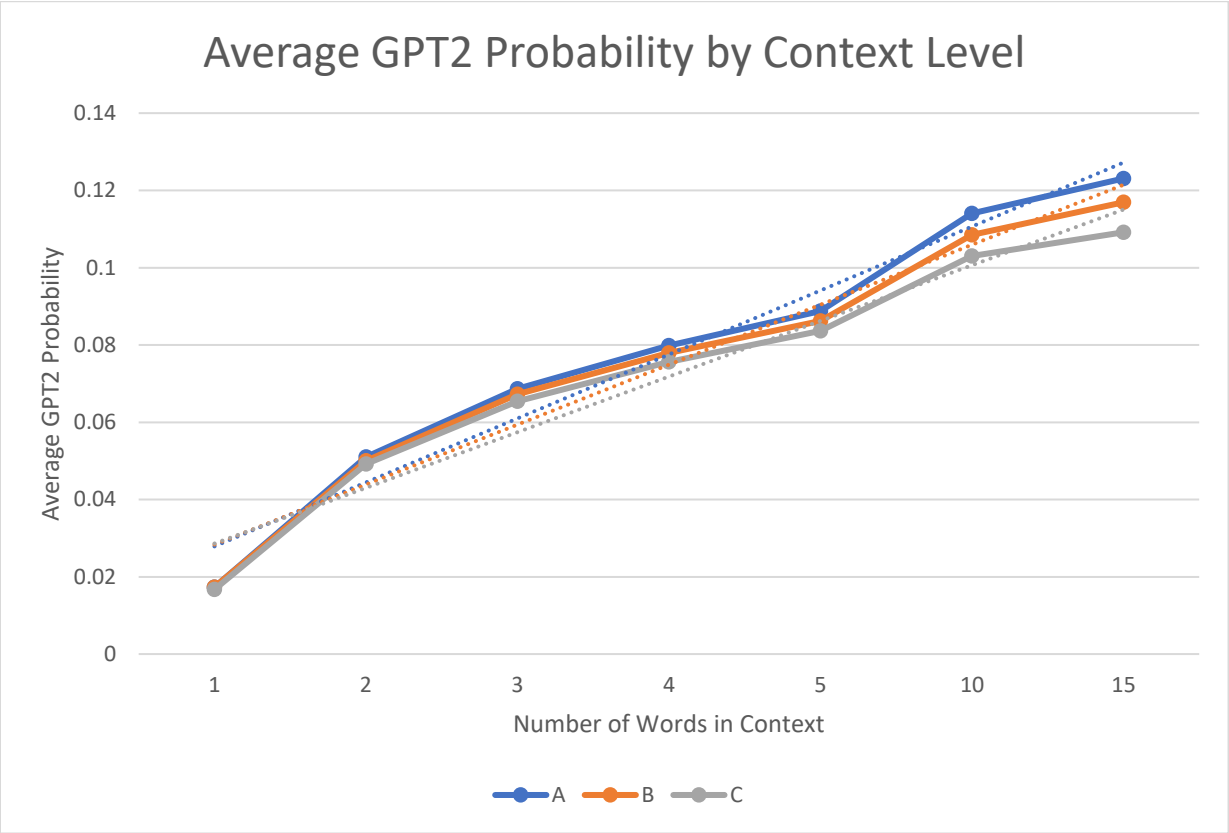
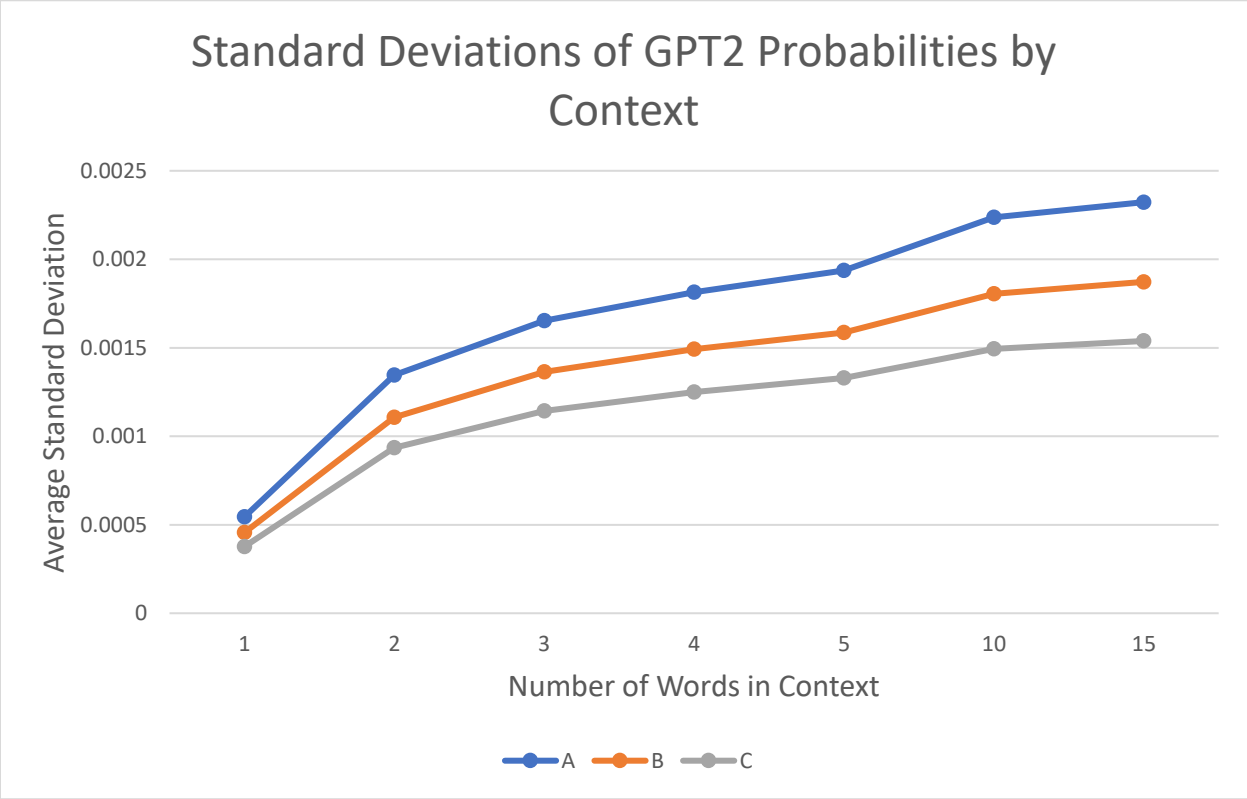


Fig 10. Average standard deviations for the GPT-2 probabilities calculated across transcripts by different amounts of context given to the model. The speech sample is from standardized interviews given to schizophrenia patients and health control. *A*: subject utterances 20 words or longer. *B*: subject utterances 15 words or longer. *C*: subject utterances 10 words or longer.



Discussion

The purpose of this study at large was to investigate the use of the language model GPT-2 as a calculator for lexical predictability that could be useful to probe mechanisms of language production. We found strong evidence that GPT-2 probabilities can be strong indicators of Cloze and trigram probabilities, while offering greater flexibility. Then, we applied GPT-2 models to naturalistic speech samples, and describe ways in which it can be a powerful tool to study theories of language production in control subjects and subjects with schizophrenia.

Corpus 1: Semi-naturalistic Vignettes

GPT-2 probability measures correlate with Cloze measures

The validity of GPT-2 as a language model used to extract predictability measures for words was strongly supported by its output's close resemblance to outputs generated by Cloze and trigram models. GPT-2 probabilities and entropies were strongly correlated with probability and entropy values extracted using the Cloze procedure (Table 1), which is considered to be the gold standard to quantify the predictability of naturalistic speech (Felice and Buttery, 2019). GPT-2 appears to be sensitive to changes in predictability throughout the course of an utterance that are detected by the Cloze procedure, and because of the somewhat consistent gap between GPT2 and Cloze probabilities, there also might be a constant that can be reliably associated with translating from one measure to the other, which would be an important contribution to translate one measure to the other (Fig 1 and Fig 2). Breaking down the entropy trends calculated by GPT-2 by each of the four different sentence positions (Fig 2), it is clear that GPT-2 has a finer

resemblance to human-characterized patterns in probability than just the overall downward trend observed over the average trend by sentence position (Fig 3), since it mimics patterns in rises and drops in Cloze entropy distributions at times by exact word positionings.

GPT-2 probability measures correlate with trigram measures

GPT-2 probabilities also mirrored patterns in probability measured by a traditional trigram model. Thus, GPT-2 appears to also be sensitive to changes in predictability throughout the course of an utterance that would be detected by the trigram procedure (Fig 4).

The GPT-2 procedure offers improvements from the Cloze and n-gram methods

GPT-2 therefore appears to be an appropriate proxy measure for trigram and Cloze probabilities, while offering substantial improvements over the other two methods. Gathering accurate Cloze probabilities from a dataset of sentences is time-consuming, expensive, and laborious given that a sufficient number of participants must be compensated for performing the task in order to have sufficient data for analysis. N-gram models, while much more accessible, are limited in vocabulary to the words in the dataset and the exact order that they appear in.

A limitation of both n-gram language models and the Cloze procedure is that there are instances where the probability of a given word can be calculated as zero, if the word in question was not part of participant responses or the n-gram training dataset. Although there are smoothing techniques that can be used to account for these probabilities, *any* zero-probability word would be treated equally, which ignores the more likely diversity in predictability even among highly unpredictable words.

Furthermore, both n-grams and Cloze as language modeling tools are locked to a specific amount of context. In the Cloze task, all participants see the same amount of context in

anticipation of the upcoming word they predict, and n-grams are limited to using *n* preceding words as context, for the entire model. GPT-2, on the other hand, can be modified at-will to make its probability calculations based on any number of words as the context. GPT-2's context flexibility can be particularly useful to studies where variations in context are relevant to predictability comparisons. This is likely the case when studying lexical prediction in schizophrenia because of its links to working memory deficits and to problems accessing global context (reviewed at length in Kuperberg, 2010), suggesting that GPT-2 would be a better language model to attain insights into the neuropsychological processes underlying schizophrenia.

Corpus 2: Natural Spoken Language

GPT-2 can be used to interpret naturalistic language production

From obtaining GPT-2 probabilities for every word in the spoken BNC, it became clear that GPT-2 predicts spoken language output very well. For both the CG-BNC and the DEM-BNC, the perplexity of GPT-2 probability calculations decreased while the mean probability increased, in every condition where the language model considered more context (Figs 5-8). These would be the patterns expected from a functional model for generating word probabilities from naturalistic speech because they indicate that the model becomes better at predicting most likely continuations as it incorporates more available information. This is an important characteristic, as it suggests GPT-2 is a viable tool for analyzing the predictability of language production in spontaneous, more naturalistic, settings.

It is worth noting that for every context category, the average GPT-2 probability was higher, and the perplexity was lower, for the CG-BNC than for the DEM-BNC. This seems to indicate that the model is better at making predictions in the CG-BNC than in the DEM-BNC. A possible explanation behind this is that the DEM model often includes transcripts with more speakers. Perhaps then, a limitation of GPT-2 probabilities is its capacity to predict the likelihood of upcoming words when more than one speaker contributes to the context. On the other hand, this could be an intrinsic characteristic of the predictability of naturalistic speech when there is convergence among different speakers (see Xu and Reitter, 2016 for discussion). A follow-up experiment that compares GPT-2 probabilities to Cloze probabilities gathered from transcripts with multiple speakers might be helpful to parse out this distinction.

A library of GPT-2 probabilities from the BNC could be used for other applications

In analyzing the use of GPT-2 to get word probabilities from the BNC, we compiled a library where every word in the spoken section of the BNC was attached to its corresponding GPT-2 probability measures across different levels of context, along with markers for sentence boundaries, changes in speakers, and changes in transcripts. This library itself could become an important tool to test different theories about the mechanisms of language production. Theories that make predictions on predictability patterns in naturalistic discourse might particularly benefit from this dataset. The organization of the library by itself would allow researchers to differentiate data by speaker, sentence, and utterance, and the additional XML flags that are associated with each of the transcripts in the CG-BNC and DEM-BNC could be used to draw further distinctions in the data, since they provide miscellaneous demographic information and information on transcript topics, since the identities of the transcripts are labeled in the library as well.

Corpus 3: Interview Transcripts

GPT-2 can be applied to distinguish patterns in probability between groups

As discussed above, GPT-2 was an accurate indicator of Cloze, the gold standard for naturalistic speech predictability, and also appears to reliably predict word probabilities in naturalistic speech transcripts. The third step then was to determine whether the GPT-2 probability paradigm can be applied to transcripts of standardized clinical interviews in control subjects.

GPT-2 probabilities were gathered over more levels of context for the transcripts of standardized interviews between an experimenter and either a schizophrenia patient or healthy control subject. The goal of analyzing transcripts in this format was to understand whether there are differences in the word predictability patterns generated over different levels of discourse between patients and healthy controls. However, it is important to note that for this, a large majority of the transcripts gathered so far were of healthy controls. The purpose of this part of the study, then, was to establish a baseline for the expected patterns of increases in predictability as context increases in a healthy population.

In all the utterances, the average GPT-2 probability calculated across all samples increased approximately linearly as the number of words GPT-2 considered in the context increased (Fig 9). The same procedure can then be used in transcripts separated by the control or schizophrenia condition in order to observe whether there are any changes in this pattern that might be indicative of changes in underlying neurocognitive processes.

We also noted that the standard deviations in the transcripts always increased as the amount of context given to GPT-2 increased (Fig 10). This is likely a reflection of the fact that as GPT-2 has more context to determine probabilities, it becomes more sensitive to highly predictable words and therefore assigns these words higher probability values whereas when the context is small, all words are unpredictable and thus there is less variance in predictability.

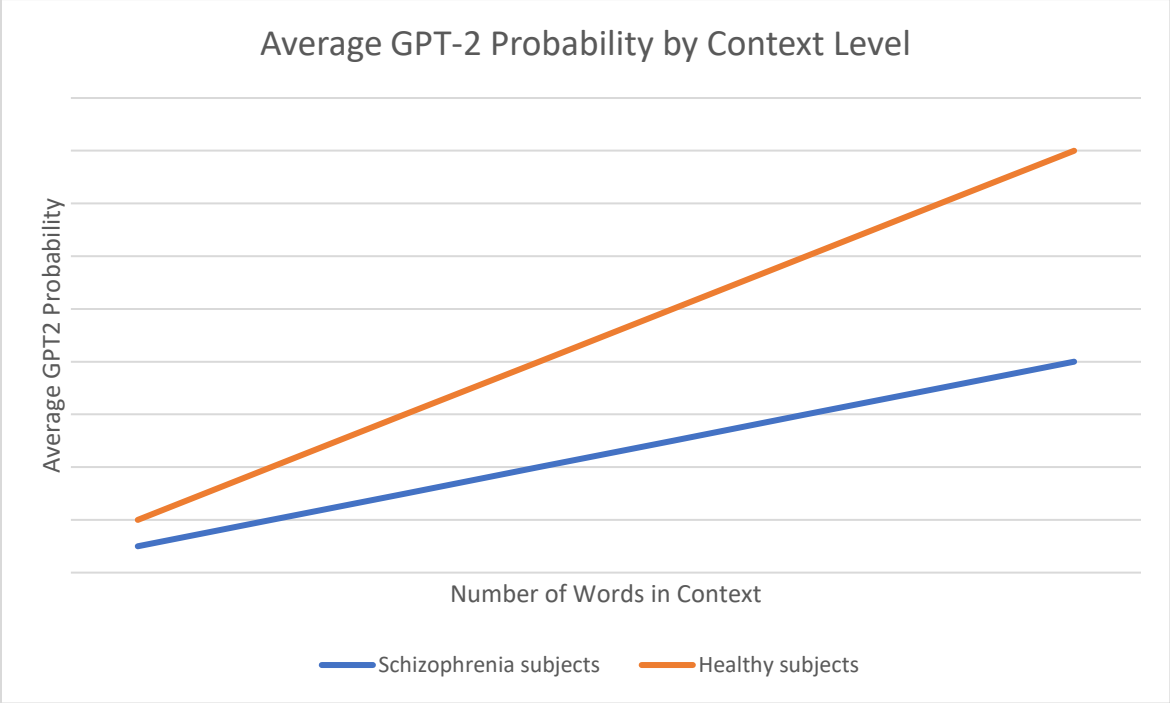
Future application: testing differences between control participants and participants with schizophrenia

We hypothesize that the greater the amount of context from participants is being considered by GPT-2, the greater the difference in word probability will be between healthy and schizophrenia subjects (see Fig 11 for illustration of this possibility). This result would provide supporting evidence that thought disorder in schizophrenia is a result of working memory deficits, where people with schizophrenia are limited in their ability to use a more global context in order to make their speech more understandable. The schizophrenia patient transcripts are currently being gathered and once they are transcribed, hypotheses about the mechanisms underlying linguistic abnormalities in neurocognitive conditions can be readily tested from speech samples.

Also, we have access to working memory task performance results associated with the participants who were interviewed. This could allow us to further test our hypothesis, since we would expect that the extent of the gap in GPT-2 probabilities between healthy and schizophrenia subjects (refer to Fig 11) will be inversely correlated with working memory performance.

In summary, we have introduced a framework that can be readily applied to study the extent to which people with schizophrenia can access more global context when producing speech. The same methodology, meaning the same modified GPT-2 script, that was applied in this last section of this study can be used on transcripts from only people with schizophrenia, and our hypothesis above will be ready to be tested as soon as we get access to more standardized transcripts of people with schizophrenia.

Fig 11. Schematic illustration of the trend in changing GPT-2 probabilities by context level that is expected to be observed from the comparison between transcripts of schizophrenia and control speech samples. *This chart does not reflect any real data thus far.*



References

Andreasen, N. C. (1979). Affective flattening and the criteria for schizophrenia. *The American Journal of Psychiatry*, *136*(7), 944–947. https://doi.org/10.1176/ajp.136.7.944

Aston, G., & Burnard, L. (1998). The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Bleuler, E. (1950). *Dementia praecox or the group of schizophrenias* (p. 548). International Universities Press.

BNC (2007). The British National Corpus, version 3 (BNC XML Edition). < http://www. natcorp.ox.ac.uk/ >

Burnard, L. (2007). Reference guide for the British National Corpus (XML Edition). Oxford University. Retrieved from http://www.natcorp.ox.ac.uk/docs/URG/ (last accessed December 2013).

Felice, M., & Buttery, P. (2019). Entropy as a Proxy for Gap Complexity in Open Cloze Tests. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 323–327. https://doi.org/10.26615/978-954-452-056-4_037

Genzel, D., & Charniak, E. (2002). Entropy Rate Constancy in Text. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 199–206. https://doi.org/10.3115/1073083.1073117

Gokaslan, A., & Cohen, V. (2019). OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2016). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *ArXiv:1511.02301 [Cs]*. http://arxiv.org/abs/1511.02301

James, F. (2000). *Modified Kneser-Ney Smoothing of n-gram Models*.

Kerns, J. G., & Berenbaum, H. (2002). Cognitive impairments associated with formal thought disorder in people with schizophrenia. *Journal of Abnormal Psychology*, *111*(2), 211–224. https://doi.org/10.1037/0021-843X.111.2.211

Kerns, J. G., Nuechterlein, K. H., Braver, T. S., & Barch, D. M. (2008). Executive Functioning Component Mechanisms and Schizophrenia. *Biological Psychiatry*, *64*(1), 26–33. https://doi.org/10.1016/j.biopsych.2008.04.027

Kuperberg, G. R. (2010). Language in Schizophrenia Part 1: An Introduction. *Language and Linguistics Compass*, *4*(8), 576–589. https://doi.org/10.1111/j.1749-818X.2010.00216.x

Lee, J., & Park, S. (2005). Working Memory Impairments in Schizophrenia: A Meta-Analysis. *Journal of Abnormal Psychology*, *114*(4), 599–611. https://doi.org/10.1037/0021-843X.114.4.599

Newby, D. (1998). "Cloze" procedure refined and modified. "Modified Cloze", "reverse Cloze" and the use of predictability as a measure of communication problems in psychosis. *The British Journal of Psychiatry: The Journal of Mental Science*, *172*, 136–141. https://doi.org/10.1192/bjp.172.2.136

Niesler, T. R., & Woodland, P. C. (1996). A variable-length category-based n-gram language model. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, *1*, 164–167 vol. 1. https://doi.org/10.1109/ICASSP.1996.540316

Qian, T., & Jaeger, T. F. (2011). Topic Shift in Efficient Discourse Production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33). https://escholarship.org/uc/item/6b0712jg

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

Salzinger, K., Pisoni, D. B., Portnoy, S., & Feldman, R. S. (1970). The immediacy hypothesis and response-produced stimuli in schizophrenic speech. *Journal of abnormal psychology*, *76*(2), 258–264. https://doi.org/10.1037/h0029887

Salzinger, K., Portnoy, S. and Feldman, R.S. (1964), VERBAL BEHAVIOR OF SCHIZOPHRENIC AND NORMAL SUBJECTS. Annals of the New York Academy of Sciences, 105: 845-860. https://doi.org/10.1111/j.1749-6632.1964.tb42967.x

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43236599*, 441–464. https://doi.org/10.2298/PSI1004441S

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sharp, B., Lubaszewski, W., & Delmonte, R. (Eds.). (2015). *Natural language processing and cognitive science: Proceedings 2015* (Prima edizione). Cafoscarina.

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. Journalism Quarterly, 30(4), 415–433.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xu, Y., & Reitter, D. (2016). Entropy Converges Between Dialogue Participants: Explanations from an Information-Theoretic Perspective. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 537–546. https://doi.org/10.18653/v1/P16-1051