A lexical predictability analysis of simulated incoherent speech

Clémentine Lévy-Fidel

1 Introduction

In this report I detail the analysis I performed to describe the effect of psychotic disorder on the lexical predictability. The goal of this analysis is to generate artificially incoherent speech with increasing levels of incoherence, or disorder level, represented as randomly shuffling an increasing proportion of words with the text samples, and to use a language model to assess how the levels of disorder affect the importance of context for predicting words in the samples accurately.

In the next Methods section I describe how I generated artificially incoherent speech data, then how I use a GPT-2 model to test the predictability of samples word by word, while varying the context length and disorder level. I also describe how I measure the relation between context length and predictability, and the behavior of this relation as a function of the level of disorder. I then discuss the results obtained in section 3, that show that the predictability decreases with increasing disorder intensity, in section 4, where I also list the limitations of my solution as well as how it could be used further.

The code of the analysis is available here.

2 Methods

2.1 Text samples generation

To simulate incoherent speech at various levels of disorder, I generated text samples with an increasing proportion of words randomly shuffled, where each of the 10 proportions ranging from 10% to 100% represents a level of disorder. The text samples were random extracts of chapters of *The Origin of Life* by Charles Darwin, found on `https://www.gutenberg.org/files/1228/1228-h/1228-h.htm`. For each disorder level, 100 samples were generated. The shuffling of the words was done as follows: a text sample was tokenized and then a proportion p of word tokens, corresponding to the disorder level considered, were sampled and their positions re-assigned to a random choice of position within them. The word tokens were finally reassembled back into one text sequence by concatenating tokens together with one whitespace between each pair of tokens.

2.1.1 Choice for tokenization

I decided to use SpaCy's tokenizer pipeline `en_core_web_sm` so as to identify individual words while handling most exceptions that a simple separation based on whitespace characters or punctuation marks (like ".", "", "-"...) would not. For example, acronyms like "H.M.S" (the Introduction chapter starts with the sentence "When on board H.M.S. 'Beagle,' ...") should be considered as a single word; contractions like "don't" should be interpreted as "do" and "n't", etc.

2.1.2 Choice of text sample length and implications on defining the boundaries of a text sample

The length of the text sample studied was mainly constrained by the model's capacity and specific tokenizing method. Indeed, GPT-2, the chosen model (see the next section), can take up to a maximum of 1'024 tokens as input, which represents about one paragraph and already shortens the randomly chosen book chapter considerably. Additionally, a text sequence given as input to the model must first be tokenized with the model's corresponding tokenizer, as the model was trained with input tokenized with its tokenizer's specific format. GPT-2's tokenizing method slightly differs from SpaCy's as it is BPE-based (Byte-Pair Encoding) , and interprets semantic a meaning in punctuation marks or whitespaces preceding a word (encoded as a `Ġ` as seen in the example below). As an example, here are how SpaCy's and GPT-2's tokenizers encode a same sequence:

- Original text sample: `When on board H.M.S. 'Beagle,' as naturalist, I was much struck with certain facts in the distribution of the inhabitants of South America, and in the`

`geological relations`

- SpaCy tokens (n=33): `When, on, board, H.M.S., ‘, Beagle, ,, ’, as, naturalist, ,, I, was, much, struck, with, certain, facts, in, the, distribution, of, the, inhabitants, of, South, America, ,, and, in, the, geological, relations`
- GPT-2 tokens (n=42) : `When, Ġon, Ġboard, ĠH, ., M, ., S, ., Ġâ, Ģ, Í, Be, agle, ,, â, Ģ, Ļ, Ġas, Ġnatural, ist, ,, ĠI, Ġwas, Ġmuch, Ġstruck, Ġwith, Ġcertain, Ġfacts, Ġin, Ġthe, Ġdistribution, Ġof, Ġthe, Ġinhabitants, Ġof, ĠSouth, ĠAmerica, ,, Ġand, Ġin, Ġthe, Ġgeological, Ġrelations`

Although the GPT-2 tokenizer is better targeted for the later use of GPT-2 in the analysis step, SpaCy tokens are more similar to the segmentation of words humans would do than GPT-2 tokens that can be parts of words (e.g. ”naturalist” giving ”natural” and ”ist”). As a consequence, shuffling GPT-2 tokens and reassembling them into a sentence later can lead to the misleading and irrelevant creation of non-existing words. Inverting first and second GPT-2 tokens in the example above for instance results in: `’ onWhen H boardM.S. .Be,agle natural as,ist was I struck much certain with in facts distribution the the of of inhabitants America South and, the in relations geological’` . As the shuffling should represent the inconsistency introduced during psychotic events in humans, it would be unlikely that the integrity of words in a sentence were not to be preserved. From this point of view, an additional control could also be to ignore punctuation marks from the candidates for shuffling (a person would surely not shuffle question marks or sentence ends), yet as in GPT-2's encoding a meaningful representation is learned from them like from other tokens, I decided to leave them as equally meaningful candidates. Finally, an important discrepancy in the use of SpaCy or GPT-2 tokens results from these different segmentation mechanisms, which leads to a different number of tokens depending on the method. Adding to the model's inherent input capacity, the issue of computation costs (see next sections) also led me to decrease the samples size below the model's maximum sequence length. I thus decided to generate samples of 64 SpaCy tokens. Finally, to maximize the diversity of samples, not only one chapter was randomly chosen for one sample, but the sample was also pooled from a randomly chosen locati within the chapter, as long as it starts as a sentence start (the first token of the sample should be either the first token of the sample or the first token succeeding a single ”.” mark).

2.1.3 Examples

Below are examples of shuffled samples for each of the disorder level.

Disorder level (%)	Effective shuffling proportion	From chapter	Original text	Shuffled text
10	0.094	3	No physiologist doubts that a stomach by being adapted to digest vegetab	general physiologist doubts that a stomach by for adapted to digest vegetable ma
20	0.188	13	The natural system is a genealogical arrangement , in which we have to dis	The natural system is a genealogical arrangement , may which we have a discove
30	0.297	4	Of this fact I will give in illustration two instances , the first which happen to	differences this in I the cases . illustration two relation , in as which happen to stan
40	0.391	7	Nor do I pretend that the foregoing remarks go to the root of the matter : n	Nor do I root that the foregoing remarks go to that is some the cases show no exp
50	0.453	1	This , again , might have been anticipated ; for the mere fact of many spec	This , again organic might inorganic been something fact any to conditions , for m
60	0.578	5	But it may be urged that when several closely - allied species inhabit the sa	: when may be urged that nearly several many intervals or allied we the same with
70	0.656	12	Thus , on the view which I hold , the natural system is genealogical in its ar	Thus different on , different which , system , them natural is the genealogical in its
80	0.750	10	As Mr. H. C. Watson has recently remarked , “ In receding from polar towa	specifically As H. earth Watson “ has the , less the naturalists value hemisphere ar
90	0.859	6	When one cell comes into contact with three other cells , which , from the s	with comes very the the frequently imitation three , surfaces has are , a same the f
100	0.969	12	A few old and intermediate parent - forms having occasionally transmitted	forms my day forms more we parent occasionally descendants the number - to th

Figure 1: Examples of shuffled text samples compared with the original text for each disorder level, as well as the effective proportion of words that have been shuffled after the process.

It is worth noticing that the effective shuffling proportion is systematically slightly smaller than the intended disorder level, attributable to the possibility that there might be redundant words within the candidates for shuffling, or that a candidate word might be randomly assigned with its initial position.

2.2 Word-by-word predictability calculation

The aim of this part was to analyze the relation between the level of disorder and the length of a word's context in the situation of assessing the predictability of that word. A word's predictability is calculated as the probability

that a model would predict this word as the next word based on the word's prior context:

$$predictability(\text{word}) = p_{model}(\text{word}_{\text{produced}} = \text{word} \mid \text{input} = \text{previous words})$$

The goal is thus to analyze the predictability of words in a text sample word by word, for different context length amplitudes, ranging from the last previous word to the maximum context available. The language model used was the small version of GPT-2 provided by *HuggingFace* in the `transformers` library [1].

In every sample, words were read and analyzed one by one. Then the predictability of each successive word was assessed repeatedly with the different context lengths to study. The average predictability by word position was then averaged disorder level by disorder level, for each context length value.

2.2.1 Input pre-processing

Tokenizing with GPT-2's tokenizer. As explained earlier, the text samples need to be tokenized with GPT-2's own tokenizer for better results. As the number of GPT-2 tokens was likely higher with GPT-2 than with SpaCy, I decided to only keep the first 64 tokens obtained after GPT-2 tokenization, despite losing a however small accuracy in the effective shuffling proportion of the text samples.

Choice of context length values. To test the effect of context length on a word's predictability, the ideal procedure would be to test a word with position i with all possible prior context ranging from the previous word (context of length 1) to all available context (context of length i), hence to test it i times, which represents $1 + 2 + ... + n = n \cdot (n + 1)/2$ times for one text sample of n tokens (here $n = 64$) and becomes computationally expensive considering the 1000 samples to test. To reduce the computational cost of studying different context length values within one sample, I decided to define a set of 6 fixed values within 1 and 62 (for a sample of 64 tokens, 63 is the last token's position, and thus 62 the largest context length possible). As increasing the context length is more likely to have a bigger effect on the model's predictions for smaller context lengths than for larger (the differences in the model's predictions are more likely to vary with 1 as compared to 3 context tokens than with 30 compared to 40), I decided to distribute the values of context length to study within the range of 1 to 63 with a logarithmic scale, so as to study "smaller" context length with more precision. The length of prior context studied are thus 1, 2, 5, 11, 27 and 62, and the values tested for a word with position i are all the values from this set that are inferior to i.

2.2.2 Next-word probability computation

Given a certain sequence as input, the model's forward pass outputs the logit probabilities of being the next word for each word in the model's dictionary. The probability distribution of words $P_{model}(\text{word}_{\text{produced}} \mid \text{input} = \text{sequence})$ is thus obtained by taking the softmax of the logits.

2.2.3 Comparison between disorder levels

Relation between lexical predictability and context length The means and standard errors are extracted from predictability scores for each disorder level and context length value. Then I fit a negative exponential function of equation $f(y) = a \cdot \exp(-b \cdot x) + c$ to the relation of the mean predictability scores as a function of context length with *SciPy*'s `polyfit` function and extracted the values of a, b and c.

Relation between the slope of lexical predictability and context length with disorder level To study the effect of the disorder level on how predictability evolves with context length, I computed the area under the fitted curve (AUC) to summarize the increase of the predictability with increasing context length as one variable, and fitted the curve representing the AUC as a function of the disorder level with an exponential curve. The AUC is calculated as: $AUC_{x0}^{x1} = [-a/b \cdot \exp(-b \cdot x) + c \cdot x]_{x=x0}^{x=x1}$.

3 Results

3.0.1 Predictability as a function of context length

The effect of context length on individual word predictability can be seen in fig. 2 for the different disorder levels. It can be seen that the predictability follows an negative exponential increase as the context length increase. Yet this behavior does not seem to apply to the highest disorder levels, where on the contrary the predictability first increases

with a negative exponential growth and then declines slightly. Finally, the predictability generally decreases with increasing disorder level, which I explore in the next part of the analysis.

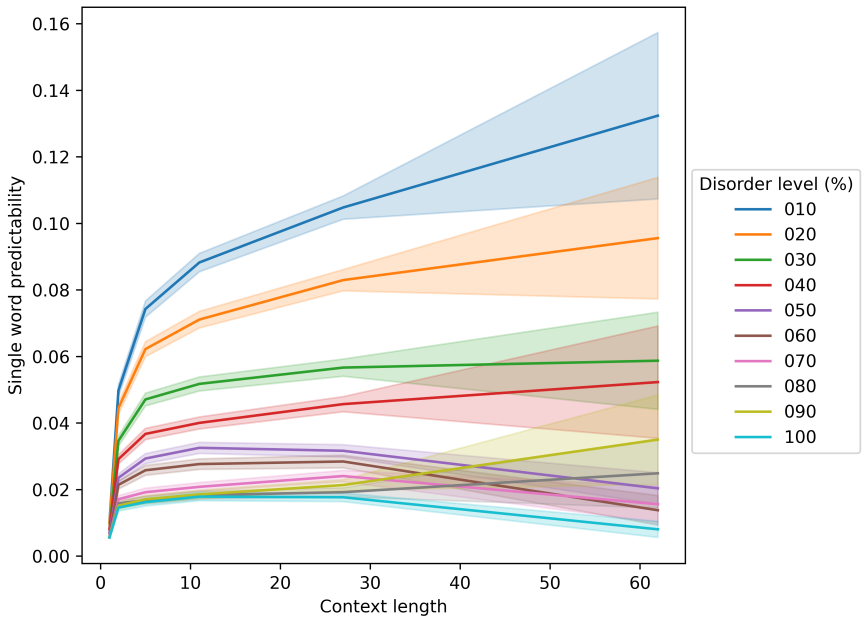

Figure 2: Curves representing the relation between word-by-word predictability as a function of the length of prior context, for the different disorder levels. The curves show the means per context length value and standard error of the mean.

3.0.2 Effect of disorder level

The curves fitting the predictability with a negative exponential are shown in fig. 3. For each fit, the R^2 coefficient representing the fraction of variance explained by the fit is displayed, and indicates that the fits is relatively good for most disorder levels, but less accurate for the levels 50 and 100.

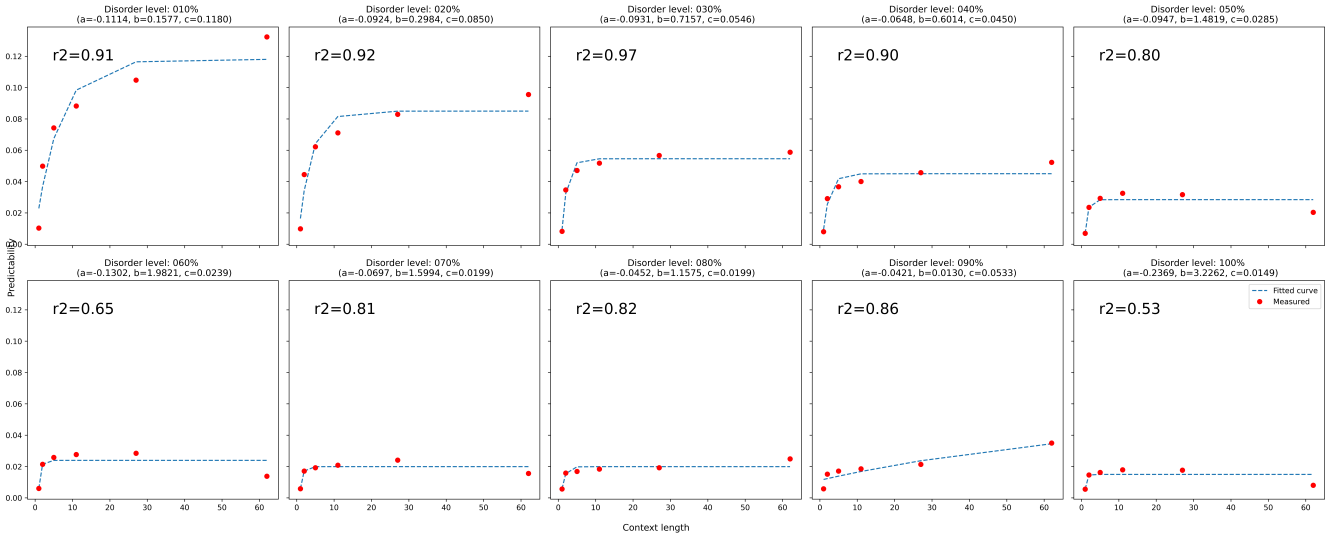

Figure 3: Predictability fit as a function of context length with an exponential function (parameters displayed above), for individual disorder levels.

The relation between the AUC and the levels of disorder is shown in fig. 4, as well as the exponential curve fit to it, that decreases rapidly at first and then slows down for higher levels of disorder. The fit's coefficient of determination is very high, indicating the the curve can be explained with the decreasing exponential behavior found.

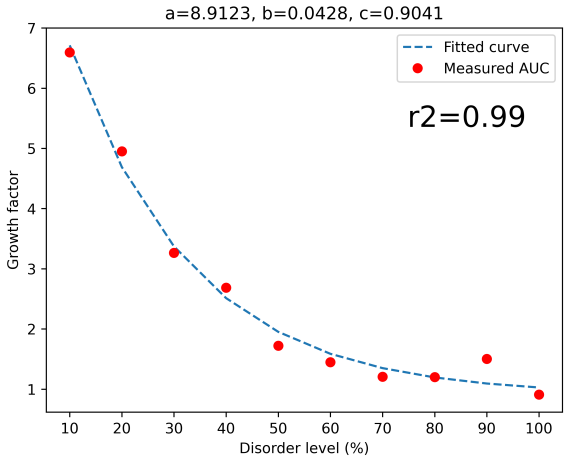

Figure 4: Predictability AUC as a function of the level of disorder, fit with an exponential function.

4 Discussion

The results allow the following observations. First, giving the model more context helps improving the predictability, and the difference is especially noticeable when the prior context is small, as suggested earlier. Increasing the level of disorder, or incoherence, in the text also implies that for a word to be predicted with a probability p, more context is needed to achieve this predictability – in other terms, the stronger the incoherence, the more context should be studied to extract meaning from a sample. Moreover, the exponential shape for the smallest levels of disorder do not seem to have achieved a very distinct plateau yet, suggesting that the predictability could still increase significantly with larger context than the ones tried in this approach. However, the intuition that predictability increases with context length follows negative exponential behavior seems to be a correct hypothesis for small disorder levels, but the falls of predictability curves for most levels above 60% could mean that for very shuffled text, adding more words in the very incoherent cases, although it allows to grasp more context information, seems to make the context's representation less and less efficient. This suggests that the amount of context alone is not sufficient for predictability, but that the context itself should be sufficiently coherent (up to 50% of incoherently positioned words), with the risk, if not, to add more uncertainty in the next-word predictions.

My solution includes several limitations. First, due to the restrictions on the computational cost I detailed in section 2, the text samples were much shorter than the maximum length allowed by the GPT-2 model, which leads to uncertainty regarding the behavior of the predictability curves in fig. 2 for longer context length (the maximal predictability score obtained is still relatively small: will the curves increase more before reaching a plateau? Will they stabilize around this plateau or will they decrease like the curves with high levels of disorder?). It would have also been more convenient to study more context length values, so as to have a better precision of the predictability at larger context length values (the hypothesis that an exponential plateau is reached in the frame studied, or will be reached, does not explain either the behaviors of the curves for high levels of disorder very well).

To conclude, this approach could be useful to predict or extract meaningful information in incoherent speech as well as to identify the level of incoherence in speech recordings, which could in return help assess the disorder intensity. However, the collapse of predictability for larger contexts in the case of stronger levels of disorder suggest that this approach would only work as long as the incoherence is not too strong, in which case the model is not able to extract meaning from the seemingly incoherent speech anymore, and worsens as more speech is given to analyze. Training a language model specifically on incoherent text samples could possibly alleviate this limitation.

5 References

[1] HF Canonical Model Maintainers. gpt2 (revision 909a290), 2022.