Uniform Complexity for Text Generation

Joseph Marvin Imperial
Human Language Technology Lab (NU HLT)
National University
Manila, Philippines
jrimperial@national-u.edu.ph

Abstract

Large pre-trained language models have shown promising results in a wide array of tasks such as narrative generation, question answering, and machine translation. Likewise, the current trend in literature has deeply focused on controlling salient properties of generated texts including sentiment, topic, and coherence to produce more human-like outputs. In this work, we introduce **Uniform Complexity for Text Generation** or **UCTG** which serves as a challenge to make existing models generate **uniformly complex** text with respect to inputs or prompts used. For example, if the reading level of an input text prompt is appropriate for low-leveled learners (ex. A2 in the CEFR), then the generated text by an NLG system should also assume this particular level for increased readability. In a controlled narrative generation task, we surveyed over 160 linguistic and cognitively-motivated features for evaluating text readability and found out that GPT-2 models and even humans struggle in preserving the linguistic complexity of input prompts used. Ultimately, we lay down potential methods and approaches which can be incorporated into the general framework of steering language models towards addressing this important challenge.

1 Introduction

In a narrative writing process, it is a general practice to maintain the complexity of text as one tries to complete the story. As such, the readability of a text depends largely on the writer's capability to reduce and simplify the structure of words and sentences (Fountas and Pinnell, 1999; DuBay, 2004). For example, when writing a book for young learners in first grade, one must take note of acceptable words that are within the vocabulary of a first grader to avoid frustration in reading that will hinder effective comprehension (Gickling and Armstrong, 1978; Guevarra, 2011). This challenge can also be emulated to natural language generators to test whether they maintain the textual complexity of prompt inputs. Figure 1 shows an ideal scenario where a prompt given by a human has the same level of reading difficulty (66.9) as the generated text by a GPT-2 model (61.9) based on the Flesch formula (Flesch, 1948). The interpretation of Flesch scores obtained means both texts can be read easily by Grade 8 and 9 learners.

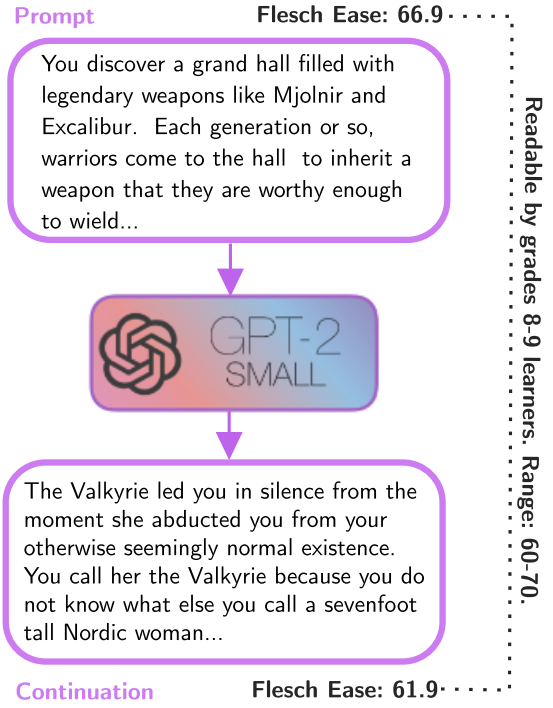

Figure 1: An illustrated example of a prompt-continuation pair produced using a GPT-2 model having the same reading level based on the Flesch formula.

Generally, the *complexity* of a narrative or any piece of text at hand can be measured by a multitude of content-based and linguistic factors. Some measures that have been explored through the years include **syntactic complexity** as equivalence of high-proficiency writing through presence of qualified syntactic phrases or *chunks* such as words

from a noun or verb phrase (Beers and Nagy, 2009; McNamara et al., 2010; Roemmele et al., 2017), **discourse complexity** by aggregating the presence of entities of a text such as mentions people, organizations, and locations which can lead to increased working memory burden (Feng et al., 2009), and **vocabulary complexity** by matching words from the text associated with a specific age-based difficulty level from developmental studies (Kuperman et al., 2012; Vajjala and Meurers, 2016) to name a few. These factors, when extracted from texts for evaluating complexity, usually follow a general linear property: the complexity of a text increases as the value of a measured feature from the text increases.

In this study, we formally introduce the task of **uniform complexity for text generation** or **UCTG** framed using an open-ended narrative generation experiment or more commonly known as story generation. To facilitate this, we used the WRITINGPROMPTS dataset from Reddit (Fan et al., 2018). To cover a wide range of measures used for approximating linguistic complexity, we extracted over 160 features drawing concepts from age-of-acquisition in developmental studies, word and phrase level part-of-speech, discourse from entity mentions, and formula-based readability indices (Lee et al., 2021). We compare text continuations from three models: humans, an off-the-shelf GPT-2 model, and a finetuned GPT-2 model. We perform a statistical test of difference for each model and for each linguistic feature. Our results show that continuations from both humans and neural language models like GPT-2, regardless of applied finetuning, generally fail to maintain the complexity levels of the prompts. We outline several potential methods that can be adapted to control the complexity of existing NLG systems. For reproducibility and transparency, we will release the processed dataset and the code upon acceptance.

2 Task Description

The proposed evaluation task can be generally adapted to any NLG framework as long as texts are both the input and output. In the case of a narrative generation setting, given a series of prompts $P = \{p_1, p_2, \ldots, p_n\}$ and their continuations $C = \{c_1, c_2, \ldots, c_n\}$ by a model M, a linguistic feature function F processes instances from both the prompt and continuation sets to get their corresponding measures $F(p_n) \in \mathbb{R}$ and $F(c_n) \in \mathbb{R}$. All calculations performed by F should be normalized with either number of word counts or a similar variable to avoid higher values from longer generated texts. This is a common practice in text complexity research to mitigate effects of length in feature values (Lu, 2012). Then, a statistical test is then applied using the two groups to measure significant difference in each linguistic complexity feature variable.

3 Generation Setup

We describe the training recipes and resources used for exploring UCTG in human and GPT-2 models.

3.1 Human Prompt and Continuation Data

For the compilation of prompts and human continuation as benchmark, we used the WRITINGPROMPTS[1] dataset collected by Fan et al. (2018). This dataset is derived from the r/WritingPrompts community of the Reddit platform where it allows users to post story premises in various context and genre for other members to *continue* with their own ideas and creativity. The current compilation is divided into train, test, and validation splits. For nature of this task, we only used the test split which contains 15,138 pairs. In addition to the data preprocessing steps done by DeLucia et al. (2021), we handpicked prompts with at least 30 words in length to ensure that the neural models will have enough context for the generation phase while capped human continuation texts with a minimum of 150 words and maximum of 300 words to avoid long-range memory limitation in the small GPT-2 model used (Sun et al., 2021). Overall, we arrive at a total of 941 prompt-human continuation pairs as benchmark for the task.

3.2 Generation Models

To test whether large neural language models are capable of UCTG, we selected the GPT-2 model (Radford et al., 2019) for analysis due to its notable extensive use in the NLP community in the narrative generation task (See et al., 2019; Xu et al., 2020; Akoury et al., 2020; Chang et al., 2021). We used two versions of GPT-2: the small, off-the-shelf version with 117M parameters from Huggingface[2] and the finetuned model from DeLucia et al. (2021) using the same WRITINGPROMPTS dataset

[1] www.reddit.com/r/WritingPrompts/
[2] https://huggingface.co/gpt2

with the train split. We no longer needed to reiterate trying out various top-p and top-k values as this has been exhaustively explored in DeLucia et al. (2021). According to the study, the best values for nucleus sampling or top-p for the narrative generation task ranges from 0.7 to 0.95 wherein generated stories are more vivid and of better quality as evaluated through human and automatic means. For this work, we set top-p to 0.7.

4 Difference in Prompt and Continuation Complexities

For the first experiment, we test for possible significant differences in the complexity values of the prompts compared with three groups: (a) the generated texts from humans of the r/WritingPrompts community, (b) the base GPT-2 model, and (c) the finetuned GPT-2 model by DeLucia et al. (2021). We used the **Welch t-test** (Welch, 1947) with Bonferroni correction resulting to an alpha level of 0.0167 (0.05/3). For the variables of interest, we extracted over 160 linguistic features using the **LingFeat tool** by Lee et al. (2021) which are often used as predictors in text readability research. The tool covers extraction of surface, lexical, phrasal, tree structure, type token, psycholinguistics, and formula-based readability features detailed in the subsections below. As mentioned in Section 2, we only used the average-based linguistic complexity features (ex. *count of noun POS / total count of words*) instead of raw counts to avoid bias or reaching extremely high values for longer sentences.

Feature	Human	GPT2	GPT2-FT
Total token x Total sent	0.0000	0.0000	0.0001
Sqrt Total token x Total sent	0.0000	0.0000	0.0001
Log token / Log sent	0.0001	0.0001	0.0001
Avr token sent	0.0001	0.0001	0.0001
Avr Syll sent	0.0001	0.0001	0.0001
Avr Syll token	0.0001	0.0001	0.0001
Avr Chars sent	0.0001	0.0001	0.0001
Avr Chars token	0.0020	0.0001	0.0001

Table 1: Shallow based features.

4.1 Shallow Features

For our first feature set, we looked at 8 shallow or surface-based textual features as shown in Table 1 such as *product and square root of total tokens by total sentences, log densities of tokens to sentences, average token count per sentence, average syllable count per sentence and per token, and average character count per sentence and token*. These linguistic features have been extensively used text complexity assessment in a wide range of languages such as in English (Flesch, 1948), French (François and Miltsakaki, 2012) and Filipino (Imperial and Ong, 2021). From the result, all of the mentioned features for the three groups appear to be significantly different with respect to the complexity of the prompt. This preempts how even humans subconsciously do not follow a uniform pattern related to surface and frequency-based properties of texts in a narrative generation task.

Feature	Human	GPT2	GPT2-FT
Flesch-Kincaid	0.0008	**0.0831**	0.0001
NARI	**0.3535**	0.0111	0.0001
Coleman-Liau	0.0001	0.0001	0.0001
SMOG	0.0000	0.0001	0.0026
Gunning-Fog	0.0001	0.0002	0.0001
Linsear	0.0001	0.0005	0.0001

Table 2: Formula based features.

4.2 Traditional Readability Features

Formula-based features for readability assessment also stem from a combinations of surface-based features such as word length, sentence length, and occurrence from a pre-defined dictionary of words. We covered 6 metrics such as *Flesch-Kincaid* (Kincaid et al., 1975), *New Automated Readability Index (NARI)* (Senter and Smith, 1967), *Coleman-Liau* (Coleman and Liau, 1975), *SMOG* (Mc Laughlin, 1969), *Gunning-Fog* (Gunning et al., 1952), and *Linsear* (Klare, 1974). Due to their ease of use, formulas such as the Flesch-Kincaid Reading Ease are integrated in most text-based applications. From the results in Table 2, majority of the narrative continuations from humans, base GPT-2 model, and finetuned GPT-2 model were significantly different except for two: the Flesch-Kincaid score obtained by the baseline GPT-2 model and the NARI score by humans. The formula for NARI is the only one which uses number of characters as predictor which may signal that human-continued texts are sensitive to character count compared to neural model-generated texts. However, collectively, the non-uniformity of complexity levels for this group is expected as the predictors used for each formula also leverages on surface-based features as previously mentioned.

4.3 Part of Speech Features

We further the analysis by looking deeper into the text structure via part of speech (POS) tags. For this study, syntactic concepts covering *nouns, verbs,*

adverbs, adjectives, subordinating clauses and conjunction, coordinating clauses and conjunction and *prepositional phrases* were used as predictors to calculate the densities of prompts and text continuations in both sentence and token level aspect (Heilman et al., 2007; Lu, 2012). Table 3 details 47 ratio and average-based predictors which quantifies POS complexities of human and neural model text continuations. Overall, the finetuned GPT-2 model obtained the least number of complexity features that are significantly different with the prompt with 31 (16 features non-significant). This is followed by the human-generated continuations with 33 (14 features non-significant) and the baseline GPT-2 with 37 (10 features non-significant). This result suggests that fine-tuning the baseline GPT-2 model with the best value for sampling-based decoding (top-p = 0.7) somehow provides a certain level uniformity of usage of grammatical entities with respect to prompts. Looking at the overlap of non-significant features with respect to the prompt, the baseline GPT-2 model only coincided with one feature that is the *average of adjective POS per sentence*. On the other hand, the finetuned GPT-2 model obtained 7 features such as *average of coordinating conjunction POS per sentence* and *ratio of coordinating conjunction POS to adjective POS* that overlapped with complexity values of the prompt. This further supports the inference on result of finetuning that was previously mentioned.

4.4 Type Token Features

Aside from looking at the average or ratio-based densities of POS tags locally, syntactic complexity can also be measured via densities of collective (more than one) POS tags per sentence. Table 4 details 5 type-token ratio (TTR) based measures: *simple type-token ratio* (O'Loughlin, 1995) , *correlated type-token ratio* (Carroll, 1964), *bilogarithmic type-token ratio* (Herdan, 1960), *Uber index* (Dugast, 1978), and *simple lexical diversity*. Type token features provide a quantified measure of unique word types (ex. combination of nouns, verbs, adjectives, and adverbs) normalized by the total number of words in a segment of a language. The variations of TTR have been studied over the years to minimize effects of sentence length when calculating the values (Herdan, 1960; Tweedie and Baayen, 1998). From the results, the human-continued text measured by the Uber index, a controlled metric for lexical diversity, was the only non-significant feature. These results may suggest that texts generated by humans, base GPT-2, and finetuned GPT-2 model may assume notable difference in densities of various basic grammatical components such as nouns, verbs, adjectives, and adverbs with respect to prompts used.

Feature	Human	GPT2	GPT2-FT
Avr Noun POS sent	0.0001	**0.0194**	0.0001
Avr Noun POS token	0.0001	0.0001	0.0001
Noun POS to Adj POS	0.0002	**0.1467**	0.0025
Noun POS to Verb POS	0.0001	0.0001	0.0001
Noun POS to Advrb POS	0.0001	0.0001	**0.3215**
Noun POS to SubrdConj	0.0001	0.0001	0.0096
Noun POS to CordConj	0.0001	**0.0814**	0.0002
Avr Verb POS sent	**0.1531**	0.0001	0.0001
Avr Verb POS token	0.0001	**0.6279**	0.0001
Verb POS to Adj POS	**0.0244**	0.0006	0.0001
Verb POS to Noun POS	0.0001	0.0001	0.0001
Verb POS to Advrb POS	**0.2302**	0.0001	0.0001
Verb POS to SubrdConj	0.0001	0.0001	0.0001
Verb POS to CordConj	0.0001	0.0001	0.0001
Avr Adj POS sent	**0.0557**	**0.0735**	0.0001
Avr Adj POS token	**0.7609**	0.0013	**0.6105**
Adj POS to Noun POS	**0.1982**	0.0027	0.0001
Adj POS to Verb POS	0.0001	0.0092	0.0024
Adj POS to Advrb POS	0.0001	0.0001	**0.0506**
Adj POS to SubrdConj	0.0001	0.0001	**0.0408**
Adj POS to CordConj	0.0001	0.0011	**0.5120**
Avr Advrb POS sent	0.0001	**0.0411**	0.0001
Avr Advrb POS token	0.0001	**0.0668**	**0.3566**
Advrb POS to Adj POS	0.0001	**0.0726**	**0.0429**
Advrb POS to Noun POS	0.0001	0.0008	0.0001
Advrb POS to Verb POS	0.0001	0.0047	0.0001
Advrb POS to SubrdConj	0.0001	0.0001	0.0138
Advrb POS to CordCobj	0.0001	**0.0247**	**0.5228**
Avr SubrdConj sent	**0.2546**	0.0001	0.0001
Avr SubrdConj token	**0.4488**	0.0001	0.0023
SubrdConj POS to Adj POS	**0.1985**	0.0001	**0.2602**
SubrdConj POS to Noun POS	**0.2955**	0.0001	0.0024
SubrdConj POS to Verb POS	0.0002	0.0001	0.0001
SubrdConj POS to Advrb POS	0.0002	0.0001	**0.8923**
SubrdConj POS to CordConj POS	0.0001	0.0001	**0.2257**
Avr CordConj POS sent	0.0001	0.0001	0.0001
Avr CordConj POS token	**0.2407**	0.0001	**0.4773**
CordConj POS to Adj POS	**0.0208**	0.0001	**0.1113**
CordConj POS to Noun POS	**0.2194**	0.0001	0.0037
CordConj POS to Verb POS	0.0001	0.0006	0.0001
CordConj POS to Advrb POS	0.0004	0.0001	**0.0220**
CordConj POS to SubrdConj POS	0.0001	0.0001	**0.0457**
Avr Content Words sent	0.0017	**0.6532**	0.0001
Avr Content Words token	0.0001	0.0001	**0.4892**
Avr Function Words token	**0.0198**	0.0001	0.0001
Avr Function Words token	0.0001	0.0001	0.0001
Content to Function Words	0.0001	0.0001	0.0001

Table 3: Part of speech based features.

Feature	Human	GPT2	GPT2-FT
Simple TTR	0.0001	0.0000	0.0001
Correlated TTR	0.0000	0.0109	0.0001
BiLogarithmic TTR	0.0001	0.0001	0.0001
Uber Index	**0.6039**	0.0001	0.0001
Lexical Diversity	0.0001	0.0001	0.0001

Table 4: Type token based features.

4.5 Lexical Variation Features

In complement to calculating ratios and averages of word-level POS complexities, lexical variation can also signal difficulty via densities of unique

grammatical components (Lu, 2012). Table 6 describes 12 lexical variation-based features focusing on *simple, squared,* and *corrected versions* of unique counts of POS such as *nouns, verbs, adjectives* and *adverbs* normalized its total in a sentence. From the results, only the *simple unique adverb count per sentence* from the human generated continuations and the *corrected unique verb variation count per sentence* from the baseline GPT-2 model obtained unsiginicant results. This finding may suggest that the densities of unique POS tags is completely different with the prompt values regardless of any human or neural-based generation mechanism.

Feature	Human	GPT2	GPT2-FT
Avr Noun phrs sent	**0.0280**	0.0001	0.0001
Avr Noun phrs token	0.0001	0.0121	0.0001
Noun phrs to Verb phrs	0.0001	0.0001	0.0001
Noun phrs to SubClaus	0.0001	**0.0757**	**0.0222**
Noun phrs to Prep phrs	0.0047	0.0067	0.0001
Noun phrs to Adj phrs	0.0001	0.0001	0.0001
Noun phrs to Adv phrs	0.0001	0.0001	0.0001
Avr Verb phrs sent	**0.2050**	0.0001	0.0008
Avr Verb phrs token	0.0001	0.0001	0.0001
Verb phrs to Noun phrs	**0.0335**	0.0001	0.0001
Verb phrs to SubClaus	0.0001	0.0001	0.0001
Verb phrs to Prep phrs	**0.5188**	**0.0168**	0.0001
Verb phrs to Adj phrs	0.0001	0.0001	0.0001
Verb phrs to Adv phrs	0.0001	0.0001	0.0001
Avr SubClaus sent	**0.7199**	0.0001	0.0021
Avr SubClaus token	0.0003	0.0001	0.0001
SubClaus to Noun phrs	**0.2813**	0.0001	0.0001
SubClaus to Verb phrs	0.0033	**0.8824**	0.0001
SubClaus to Prep phrs	**0.0169**	**0.2844**	0.0001
SubClaus to Adj phrs	0.0001	0.0001	0.0001
SubClaus to Adv phrs	0.0001	0.0001	0.0001
Avr Prep phrs sent	**0.1307**	0.0001	0.0001
Avr Prep phrs token	**0.2976**	**0.6734**	0.0001
Prep phrs to Noun phrs	**0.1831**	**0.5031**	0.0001
Prep phrs to Verb phrs	0.0001	0.0001	0.0001
Prep phrs to SubClaus	0.0001	**0.0288**	0.0014
Prep phrs to Adj phrs	0.0001	0.0001	**0.1481**
Prep phrs to Adv phrs	0.0001	0.0001	**0.1013**
Avr Adj phrs sent	0.0001	0.0029	**0.2439**
Avr Adj phrs token	0.0001	0.0001	0.0001
Adj phrs to Noun phrs	0.0002	0.0075	0.0001
Adj phrs to Verb phrs	**0.3468**	**0.2768**	**0.4599**
Adj phrs to SubClaus	0.0001	0.0014	0.0001
Adj phrs to Prep phrs	**0.6291**	**0.7919**	0.0001
Adj phrs to Adv phrs	0.0001	0.0001	0.0001
Avr Adv phrs sent	0.0001	0.0003	0.0001
Avr Adv phrs token	0.0001	**0.3127**	0.0001
Adv phrs to Noun phrs	0.0001	**0.9194**	**0.4361**
Adv phrs to Verb phrs	0.0001	0.0001	0.0001
Adv phrs to SubClaus	0.0001	**0.7101**	**0.2759**
Adv phrs to Prep phrs	0.0001	**0.0603**	0.0001
Adv phrs to Adj phrs	0.0001	0.0001	0.0001

Table 5: Phrasal based features.

4.6 Phrasal Features

Moving on to longer sequences of part-of-speech complexities, we also measure phrase-level linguistic features. Table 5 shows 42 phrase-based features centering on token and sentence ratios of grammatical components such as *noun phrases, verb phrases, adverbial phrases, adjectival phrases, subordinate phrases* and *prepositional phrases*. Overall, the human continuation obtained 30 phrasal-based features which significantly different with respect to the prompt while 29 for the baseline GPT-2 and 35 for the finetuned GPT-2 model. In contrast to the observation in word-level POS features, for phrasal-based POS, the baseline GPT-2 model have more coincided non-significant features with the prompt—6 compared to the finetuned GPT-2 model with only 3. Some of these features are ratios of *prepositional phrases to noun phrases* and *verbial phrases to prepositional phrases*. From this result, we posit that avoiding finetuning of the GPT-2 model preserves the syntactic structure of the sentence (upper portions of the tree) at phrase level but not at finegrained word level as seen in POS features at Table 3. In addition, this result also preempts why Transformer-based models are often finetuned for paraphrasing task to trigger require number of lexical swaps and syntactic diversity (Witteveen and Andrews, 2019; Krishna et al., 2020).

Feature	Human	GPT2	GPT2-FT
Simpl Noun variation	0.0004	0.0001	0.0001
Sqrd Noun variation	0.0001	0.0001	0.0001
Corr Noun variation	0.0001	0.0001	0.0001
Simpl Verb variation	0.0001	0.0001	0.0001
Sqrd Verb variation	0.0001	0.0001	0.0001
Corr Verb variation	0.0001	**0.1771**	0.0001
Simp Adj variation	0.0001	0.0001	0.0001
Sqrd Adj variation	0.0001	0.0001	0.0001
Corr Adj variation	0.0001	0.0001	0.0001
Simp Adv variation	**0.1610**	0.0001	0.0001
Sqrd Adv variation	0.0001	0.0001	0.0001
Corr Adv variation	0.0001	0.0001	0.0001

Table 6: Lexical variation based features.

4.7 Syntax Tree Features

Following the results of parse tree-based features in Table 5, analyzing the difference in parse tree depth is a unique and interesting way of measuring readability complexity as done in Schwarm and Ostendorf (2005). Table 7 describes 4 syntax tree height-based features including *average heights of regular and flattened trees* per token and sentence. From the result, there is a pattern seen with the human-generated texts as it obtained non-significance with *average regular tree height* and *average feature tree height* at a sentence-level. This suggests that neural model-based generated texts do not conform to the one property of syntax which is the parse tree height in contrast to human continuations.

Feature	Human	GPT2	GPT2-FT
Avr Tree height sent	**0.0326**	0.0001	0.0002
Avr Tree height token	0.0001	0.0001	0.0001
Avr FTree height sent	**0.1125**	0.0001	0.0001
Avr Ftree height token	0.0001	0.0001	0.0001

Table 7: Syntax tree based features.

4.8 Psycholinguistic Features

We also look at psycholinguistic variables in reading using external wordlists such as the Age-of-Acquisition (AOA) database compiled by Kuperman et al. (2012) which contains over 30,000 English content words and the SubtlexUS by (Brysbaert and New, 2009) which is a compilation of over 74,000 word forms with frequency values extracted from 8,000 general films and series. These special databases contain words that are associated to various age levels where children expected to learn when they reach the stage. The works of Vajjala and Meurers (2016) and Chen and Meurers (2016) both have leveraged on these predictors in readability assessment and text familiarity. Using the LingFeat tool, we extract over 26 *token, lemma, and sentence-based normalizations of AOA and SubtlexUS variations*. Results from Table 8 show that not a single model has coincided with each other in terms on non-significant feature. The baseline GPT-2 model's continuations are significantly different from the prompts with respect to psycholinguistic features used while the fine-tuned GPT-2 model only obtained non-significance from the averages of token-based features from SubltexUS. This may faintly suggest that the fine-tuned GPT-2 model simply reuse some of the words present from the prompt that are recognized by or included in the SubtlexUS database.

4.9 Discourse Features

For the last linguistic feature set investigated, we look at discourse in the form of *averages of unique and non-unique entity presence* that can affect working memory load as well as *local coherence distance measures* which captures distribution and transitions of entities in a passage (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013). Feng et al. (2009) previously applied these cognitively-motivated features for assessing reading difficulty in the case of adults with intellectual disabilities. Table 9 shows 10 discourse-level features extracted from the prompt-continuation pairs where majority of the features have shown significant difference

Feature	Human	GPT2	GPT2-FT
AOA word sent	0.0001	0.0001	0.0001
AOA word token	0.0001	0.0001	0.0001
AOA lemma sent	0.0001	0.0001	0.0001
AOA lemma token	0.0001	0.0001	0.0001
AOA lemma Bird sent	0.0001	0.0001	0.0001
AOA lemma Bird token	0.0094	0.0001	0.0001
AOA Bristol sent	0.0001	0.0042	0.0001
AOA Bristol token	0.0001	0.0001	0.0001
AOA CortKhanna sent	0.0001	0.0042	0.0001
AOA CortKhanna token	0.0001	0.0001	0.0001
SubtlexUS sent	0.0001	0.0001	0.0001
SubtlexUS token	0.0001	0.0001	**0.6334**
SubtlexUS CD sent	0.0002	0.0001	0.0001
SubtlexUS CD token	0.0001	0.0001	0.0001
SubtlexUS FREQ sent	0.0001	0.0001	0.0001
SubtlexUS FREQ token	0.0001	0.0001	**0.0182**
SubtlexUS CDL sent	0.0001	0.0001	0.0001
SubtlexUS CDL token	0.0001	0.0001	0.0001
SubtlexUS SUBTL sent	0.0001	0.0001	0.0001
SubtlexUS SUBTL token	0.0001	0.0001	**0.6334**
SubtlexUS Lg10WF sent	0.0001	0.0001	0.0001
SubtlexUS Lg10WF token	**0.8759**	0.0001	0.0001
SubtlexUS SubLCD sent	0.0002	0.0001	0.0001
SubtlexUS SubLCD token	0.0001	0.0001	0.0001
SubtlexUS LgCD sent	0.0001	0.0001	0.0001
SubtlexUS LgCD token	0.0003	0.0001	0.0001

Table 8: Psycholinguistics based features.

for all models. This finding may hint that the generated continuations from the three models have dissimilar levels of dependencies with respect to the prompts or vice versa.

Feature	Human	GPT2	GPT2-FT
Avr Entity sent	0.0001	**0.0881**	0.0001
Avr Entity token	0.0001	0.0001	0.0001
Avr Uniq Entity sent	0.0001	0.0001	0.0001
Avr Uniq Entity token	0.0001	0.0001	0.0001
Local Coherence PA	0.0001	0.0001	0.0087
Local Coherence PW	0.0001	0.0001	0.0087
Local Coherence PU	0.0001	0.0001	0.0001
Local Coh Dist PA	0.0001	0.0001	0.0107
Local Coh Dist PW	0.0001	0.0001	0.0107
Local Coh Dist PU	0.0001	0.0001	0.0001

Table 9: Discourse based features.

5 Correlation in Human and GPT-2 Continuations

Aside from prompt-wise comparison, we also look at which linguistic complexity feature from the GPT-2 models are correlated with the human continuations to identify. This answers the question of which model is closer to produce more *human-like* continuations in the lens of text complexity features. For this, we use **Pearson correlation** to do the continuation-wise analysis and extract the top correlated features described in Table 10. From the Table, both baseline GPT-2 and finetuned GPT-2 model obtained 7 top complexity features all from the psycholinguistics category referencing from *Age-of-Acquisition* and *SubtlexUS databases*.

GPT-2		GPT-2 FT	
SMOG	0.188	Avr Syll token	0.155
Avr Syll token	0.187	AOA Bristol token	0.102
SubtlexUS Lg10WF token	0.177	AOA CortKhanna token	0.102
SubtlexUS CD token	0.162	AOA lemma token	0.102
SubtlexUS SubLCD token	0.162	AOA word token	0.097
SubtlexUS CDL token	0.159	SubtlexUS CD token	0.093
SubtlexUS LgCD token	0.154	SubtlexUS SubLCD token	0.093
Avr Verb POS token	0.120	SubtlexUS CDL token	0.090
AOA lemma token	0.119	Avr Adv phrs token	0.080
AOA word token	0.116	Adv phrs to Noun phrs	0.079

Table 10: Top correlated complexity features of GPT-2 and finetuned GPT-2 model with human continuations.

This may suggest that even if there is a dissimilarity with the prompt-continuation comparison for all models, they are nonetheless correlated in the sense with the human continuation against other complexity features. Likewise, for both baseline and finetuned GPT-2 models, the *average syllable count per token* emerged as the top common (weakly) correlated features. However, we cannot draw solid conclusions of *human-likeness* from this specific result since all of the features are weakly correlated in the general sense (r <0.50).

6 Potential Methods for Controlling Complexity

We put forward a list of possible methods and related works of similar vein with substantial potential for enforcing UCTG in existing NLG systems which can be explored by future researchers.

Building A Uniformly Complex Prompts-Continuation Pairs Dataset. Probably the most naive approach for this task is to compile a dataset containing uniformly complex texts prompts and their corresponding continuations which belong to the same readability level or complexity spectrum. The dataset has to be large enough to finetune GPT-2 or any large pretrained neural language model used in NLG systems. This approach is similar to the work of Agrawal and Carpuat (2019) where they compiled English and Spanish news articles of diverse grade levels for sentence alignment in translation. Alternatively, one could also *split* lengthy collections of literary texts (such as ones found in Project Gutenberg[3]) assuming the resulting text splits will retain the same readability level or with range when validated. The UCTG evaluation as seen in this study can then be re-applied to the newly trained model to probe if there are non-significance from the linguistic complexity measures of continuation texts with respect to the input text prompts used.

Post-Editing with Specialized Vocabularies. Another relatively simple method with potential is applying lexical substitution of complex words from the generated continuations with words from *specialized wordlists* such as the Age-of-Acquistion (AoA) by Kuperman et al. (2012) or the Academic Vocabulary List (AVL) by Gardner and Davies (2014). These resources, often used in developmental and literacy-related studies, contain words and their corresponding normalized age values where they are typically learned by children. The updated AoA database contains over 50,000 words distributed over various categories such as nouns, verbs, and adjectives while the AVL contains 3,015 lemmas occurring in all academic domains, their word forms, inflected forms, and ratio value per million words. However, one possible limitation of this approach is that databases have limitation on the scope of the age done during collection. The AoA database, for instance, currently only captures words learnt until the age of 25. Generally, this may already suffice if ever a study would only focus on texts for early to intermediate level users.

Controlling Generations with Auxiliary Models. More advanced techniques for control is through anchoring the complexity features of text with existing *plug-and-play* methods (Dathathri et al., 2019; Pascual et al., 2021) leveraging on attribute models. These attribute models, commonly in the form of simple bag-of-words model or a linear classifier, operate by *shifting* the distribution of the generated text continuation in a way that it will gradually conform to the

[3]https://www.gutenberg.org/

desired attribute by acting as a *scorer* of quality of texts generated. For UCTG, the desired attribute model can be the target readability level or complexity measure which is similar to the prompt. In literature, attribute models can be stacked together to improve generation results. Thus, a readability attribute model can be developed and combined with attribute models of fluency, topic, and sentiment for better quality of results. One can also rethink the way of gauging the complexity of texts for generation tasks. Recently, August et al. (2022) explored controllable generation of scientific definitions using a novel re-ranking feature. In this work, a BART model was prompted to generated 100s of candidate definitions which were *ranked* by quality using a linear SVM or BERT-based discriminator.

7 Moving Forward

We propose a new challenge towards making existing natural language generation systems controllable by taking a perspective in text complexity. While previous works have explored only a small portion of linguistic complexity features (Roemmele et al., 2017; See et al., 2019), we provide the bigger, more in-depth picture by analyzing over a hundred linguistic complexity features to show that even human continuations, often regarded as gold-standard in literature, generally do not maintain uniformity with the distributional linguistic properties of the prompts. In hindsight, for us humans, we do not have the capability to measure text complexity values while reading on-the-fly which makes our judgment unreliable and subjective (Deutsch et al., 2020). Thus, in the field of education where proper matching of reading materials is crucial (DuBay, 2004), this raises the need for a tool that can act as writing assistants for authors and educators towards making children's literature fit for various levels of audiences and conforms to specially-defined metrics such as the Common European Framework of Reference for Languages (CEFR). Similarly, the continuations from the GPT-2 models are no better when compared to human results showing no significant correlation with any of the linguistic features investigated. To this end, we highlight possible future directions such exploration on compilation of uniformly complex prompt-continuation pairs, post-editing with specialized wordlists, and exploiting external models as discriminators. We foresee increased accessibility of powerful NLG systems if we also consider the angle of readability and text complexity whenever we build upon these resources for public use.

8 Limitations

This study tests whether humans and GPT-2, often used as a baseline model for NLG tasks, exhibit any solid and significant evidence of being able to maintain stable linguistic complexity measures. From a wide range of structured and non-structured NLG tasks, we decide to restrict our experiments to a narrative generation task as this is more closely related to the story writing process often done in the education field where materials are strictly classified into various age or grade level categories.

9 Ethical Considerations

We foresee no serious ethical concerns for this evaluation study.

10 Acknowledgments

The authors would like to thank Alexandra DeLucia from the Johns Hopkins University Center for Language and Speech Processing for their help in providing the small version of the finetuned GPT-2 model. This project is funded by the National University - Center for Research and the Google AI Tensorflow Faculty Award of main proponent.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

Tal August, Katharina Reinecke, and Noah Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317,

Dublin, Ireland. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Scott F Beers and William E Nagy. 2009. Syntactic complexity as a predictor of adolescent writing quality: Which measures? which genre? *Reading and Writing*, 22(2):185–200.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

John B Carroll. 1964. Language and thought. *Foundations of Modern Psychology Series*.

Haw-Shiuan Chang, Jiaming Yuan, Mohit Iyyer, and Andrew McCallum. 2021. Changing the mind of transformers for topically-controllable language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2601–2611, Online. Association for Computational Linguistics.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94, San Diego, CA. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2021. Decoding methods for neural narrative generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 166–185, Online. Association for Computational Linguistics.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

William H DuBay. 2004. The principles of readability. *Online Submission*.

D Dugast. 1978. *Sur quoi se fonde la notion d'étendue théoretique du vocabulaire?*, volume 46.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237, Athens, Greece. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Irene C Fountas and Gay Su Pinnell. 1999. *Matching Books to Readers: Using Leveled Books in Graded Reading, K-3*. ERIC.

Thomas François and Eleni Miltsakaki. 2012. Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57, Montréal, Canada. Association for Computational Linguistics.

Dee Gardner and Mark Davies. 2014. A new academic vocabulary list. *Applied Linguistics*, 35(3):305–327.

Edward E Gickling and David L Armstrong. 1978. Levels of instructional difficulty as related to on-task behavior, task completion, and comprehension. *Journal of Learning Disabilities*, 11(9):559–566.

Rowena C. Guevarra. 2011. Development of a Filipino text readability index.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

Robert Gunning et al. 1952. Technique of clear writing.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467, Rochester, New York. Association for Computational Linguistics.

Gustav Herdan. 1960. *Type-token Mathematics*. Mouton.

Joseph Marvin Imperial and Ethel Ong. 2021. Diverse linguistic features for assessing reading difficulty of educational filipino texts. *arXiv preprint arXiv:2108.00241*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

George R Klare. 1974. Assessing readability. *Reading Research Quarterly*, pages 62–102.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication*, 27(1):57–86.

Kieran O'Loughlin. 1995. Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, 12(2):217–237.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Melissa Roemmele, Andrew S Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *SIGKDD 2017 Workshop on Machine Learning for Creativity*, pages 13–17.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China. Association for Computational Linguistics.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *arXiv preprint arXiv:1603.06009*.

Bernard L Welch. 1947. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong. Association for Computational Linguistics.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.