Harnessing GPT to Study Second Language Learner Essays: Can We Use Perplexity to Determine Linguistic Competence?

Ricardo Muñoz Sánchez[†], **Simon Dobnik**[‡], **Elena Volodina**[†]

[†] Språkbanken Text, University of Gothenburg, Sweden

[‡] CLASP, FLoV, University of Gothenburg, Sweden

{ricardo.munoz.sanchez,simon.dobnik,elena.volodina}@gu.se

Abstract

Generative language models have been used to study a wide variety of phenomena in NLP. This allows us to better understand the linguistic capabilities of those models and to better analyse the texts that we are working with. However, these studies have mainly focused on text generated by L1 speakers of English. In this paper we study whether linguistic competence of L2 learners of Swedish (through their performance on essay tasks) correlates with the perplexity of a decoder-only model (GPT-SW3). We run two sets of experiments, doing both quantitative and qualitative analyses for each of them. In the first one, we analyse the perplexities of the essays and compare them with the CEFR level of the essays, both from an essay-wide level and from a token level. In our second experiment, we compare the perplexity of an L2 learner essay with a normalised version of it. We find that the perplexity of essays tends to be lower for higher CEFR levels and that normalised essays have a lower perplexity than the original versions. Moreover, we find that different factors can lead to spikes in perplexity, not all of them being related to L2 learner language.

1 Introduction

In the past couple of years we have seen a fast development in the capabilities of decoder-only language models, such as GPT-4 (OpenAI et al., 2024), LLaMA (Touvron et al., 2023), and BLOOM.[1] These models have been increasingly deployed in a wide variety of applications such as machine translation (Qian, 2023) and financial (Li et al., 2023) and legal applications (Kwak et al., 2023). In the context of second language (L2) educational applications, these models have been deployed to different subtasks, with varying degrees of success (Naismith et al., 2023; Yancey et al., 2023).

Even though they excel at a multitude of NLP tasks, the inner workings of large language models are obscure. This means that it is complicated, if not impossible, to verify that a model has learned actual linguistic features instead of making spurious correlations. The same issue is true when attempting to determine how the model arrived at specific decisions (Blevins et al., 2023). This can be tricky, especially in high-stakes situations such as educational applications.

One such example is the evaluation and assessment of second language performance, as the result of such assessment can alter the life opportunities a person has access to (education, job offers, among others). When dealing with text, we want to be able to understand how systems interact with text from second language learners. This would allow us to properly build models for tasks such as second language assessment, for grammatical error correction, among others, while complying with the right to an explanation (Official Journal, 2016).

One way to do so is to analyse how much text diverges from what a language model expects. This can be done using perplexity, a statistical concept that measures the probability of a sequence given an estimator and has been interpreted in an intuitive manner as a measure of "surprisal" (Dobnik et al., 2018; Niu and Penn, 2020). However, it has been mostly used to study texts of first language speakers of English. To address this gap, we aim to study how perplexity interacts with texts generated by second language learners of Swedish.

We hypothesise that the perplexity of a decoder-only model is related to the complexity of the text in an L2 speaker's essay. In this paper we aim to answer the following research questions:

- To what degree can the linguistic competence of a learner (as evidenced in an essay) be estimated using perplexity from language models?

[1] https://bigscience.notion.site/BLOOM-BigScience-176B-Model-ad073ca07cdf479398d5f95d88e218c4

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 414–427

June 20, 2024 ©2024 Association for Computational Linguistics

- To what degree does the perplexity of the language model correspond to CEFR[2] levels?

We have used GPT-SW3 (Ekgren et al., 2023) as our language model for this study. It is based on the GPT series (Radford et al., 2019; Brown et al., 2020) and trained on data of several Nordic languages. We give more details about it in Section 3.1. For the L2 learner essays we used Swell (Volodina et al., 2016a; Volodina, 2024), a collection of corpora of L2 learner essays in Swedish. A more in-depth explanation of its contents and how the essays were collected can be found in Section 3.2. We describe perplexity and some of its intuitive interpretations in Section 3.3.

We ran two sets of experiments. In Section 4 we show the perplexity of the essays and see how it is distributed both across levels and within the essays. This analysis is done both in a statistical manner in Section 4.1 and in a linguistic manner in Section 4.2. The second set of experiments is a comparison between the perplexity from original essays written by L2 students and normalised versions of these essays, described in Section 5.

2 Related Work

2.1 NLP for Second Language Learning

There have been several ways in which NLP has been used for second language learning. The two most relevant for us are automated essay scoring and grammatical error correction.

Automated essay scoring (AES) of L2 learner texts is a task in which we have a system that takes a text generated by an L2 learner and assigns a grade or level to it. This can be done using CEFR levels, but different levels or scales have also been used in the past. Despite their ubiquity in NLP and machine learning in general, deep learning had not been used in AES until 2016 (Alikaniotis et al., 2016; Taghipour and Ng, 2016). Even though there have been more works that use deep learning for this task, Mayfield and Black (2020) warn that their performance might not be good enough to justify the lack of transparency and the increased computational costs. As for decoder-only models, they were first used for this task in 2023, with mixed results (Naismith et al., 2023; Yancey et al., 2023).

[2]CEFR stands for Common European Framework of Reference for Languages. It is a framework to evaluate foreign language learning and assigns one of six reference levels to determine the proficiency level of a second language speaker (Council of Europe, 2001).

Grammatical error correction (GEC) is a task in which we have a system that takes a text assumed to have some sort of error or non-standard language and returns a normalised version of the same text. It is important to note that, despite the name of the task, the errors in the original (or source) text are not limited to grammar and often include other kinds or errors, such as lexical, orthographic, among others (Bryant et al., 2017). It is often seen as variation of machine translation, with the source language being the non-normalised language and the target being the normalised one (Wang et al., 2021). Because of this, sequence-to-sequence neural models have often been used for this task, including decoder-only models (e.g. Flachs et al., 2019).

Most of the work done so far in this area has focused on English. However, the advances for Swedish are scarce, despite it being a language with relatively good language technology resources. The Swell corpus collection (Volodina, 2024) contains corpora both for AES (Swell-pilot) and for GEC (Swell-gold). As far as we are aware, the current state of the art of AES in Swedish is that of Pilán et al. (2016) and Volodina et al. (2016b). They use a feature-based approach using length-based, lexical, morphological, syntactic, and semantic features. In terms of GEC, the most recent approach is that by Kurfalı and Östling (2023), who used a transformer-based model.

2.2 Language models as Predictors of Grammaticality

As Lappin (2023) points out, the discussion of linguistic capabilities of large language models ranges from (overstated) claims of their sentience and the arrival of artificial general intelligence to skepticism and dismissal. Because of that, he argues it is is essential to explore the capabilities of these models. One way that this has been done is by evaluating how much texts generated by humans diverge from what a language model expects.

One possible approach is by evaluating the grammaticality or linguistic acceptability of a text. The idea is to give a system a text that it has to determine whether it is grammatically correct or not. There are two main approaches through which this can be done. The first one is as a classification task, assigning each sentence a class that determines whether a sentence is grammatically correct or not (Klezl et al., 2022).

Another approach is by checking whether a text is likely to appear in text generated by a language model or not (Lau et al., 2017). In particular, perplexity has been used as a way to determine how much a model expects the tokens within a text to appear (Niu and Penn, 2020). It has subsequently been interpreted as a measure of "surprisal".

3 Methodology and Experimental Settings

3.1 GPT-SW3

Our objective is to determine how much L2 Swedish learner's texts differ from what a generative language model would expect.

In order to do this, we use GPT-SW3 (Ekgren et al., 2023), an auto-regressive model based on the GPT series of models (Radford et al., 2019; Brown et al., 2020). It was trained on a large dataset called The Nordic Pile (Öhman et al., 2023), a 1.3TB dataset containing large dumps of several websites in the Nordic languages.[3]

We decided to use this model as it is to our knowledge the largest and best performing generative model currently available for the Swedish language. Our assumption is that it will be able to model Swedish in a similar way to how L1 speakers write across a variety of domains, thus allowing it to identify when an L2 speaker's sentences differ from what an L1 speaker would write.

3.2 Dataset

To compare how GPT-SW3 works for different CEFR levels of language learner essays, we have used the Swell corpus collection (Volodina, 2024). It is divided into two corpora, Swell-pilot (Volodina et al., 2016a) and Swell-gold (Volodina et al., 2019).

Swell-pilot consists of 502 essays divided into three sub-corpora, collected between 2012 and 2016. All essays are anonymised and annotated for CEFR level. However, there are six essays that lack a level, so we have ignored them for the purposes of this paper.

Swell-gold consists of 502 essays that were collected between 2017 and 2021. They are pseudonymised and include both the original version and a normalised version of the essays. They also contain level indications, which, however, do not align with the CEFR levels. These levels were determined by using the course that the student was taking as a proxy for proficiency, not by an actual analysis of learner performance.

Level	N° of Essays
A1	59
A2	143
B1	86
B2	105
C1	96
C2	7 (Swell-pilot) 43 (Swell-gold)

Table 1: Distribution of the CEFR levels in Swell-pilot. Note that we added extra essays from Swell-gold to have a more representative sample of the C2 level.

Level	N° of Essays
Beginner (*Nybörjare*)	289
Intermediate (*Fortsättning*)	45
Advanced (*Avancerad*)	168

Table 2: Distribution of the proficiency levels in Swell-gold. The text in parenthesis is the name for the level used in the metadata (in Swedish).

In our first experiment (Section 4) we use all the essays from Swell-pilot that have a CEFR level assigned to them, as they showcase a good distribution from the different levels, as seen from Table 1. The only exception is the C2 level which only has seven essays. To make up for this, we randomly sampled 43 of the normalised version of the essays in Swell-gold by advanced speakers as we assumed that it would more closely resemble those by C2 second language learners.

In our second experiment (see Section 5) we use both the original and the normalised versions of all of the Swell-gold essays. The distributions of the proficiency annotations of the essays can be found in Table 2.

3.3 Perplexity as a Measure of Surprisal

Perplexity is the probability that an observation is made by an estimator. When dealing with generative models, this is the probability that a sequence S appears in a natural language L. When we use a language model M, we do so as it approximates (or models) language L. Thus, we can intuitively interpret the perplexity PP_M as a way to measure how "surprised" the model M is to see sequence S.

[3]The languages included are Danish, Faroese, Icelandic, Norwegian, and Swedish. For more information about the contents of the dataset, read AI Sweden's blog post: https://medium.com/ai-sweden/the-nordic-pile-a8d5aaf3db60

Now, perplexity is defined in mathematical terms as the probability that an estimator (in this case M) sees an observation S (Jelinek et al., 2005). The best way to calculate this for a generative model is by taking the product of the probabilities of a token given the previously generated ones:

$$PP_M(S) = \mathbb{P}(S)^{-|S|} = \prod_{i \leq |S|} \mathbb{P}(S_i|S_{<i})^{-|S|}$$

where S_i denotes the i-th token of S and $S_{<i}$ the sequence S up to S_i.

Given that this is a very small number, we risk having an underflow in our calculations[4]. Because of this, we are better off using the log-likelihood as opposed to using the regular likelihood. Thus, we have

$$\begin{aligned}\log PP_M(S) &= \log \prod_{i \leq |S|} \mathbb{P}(S_i|S_{<i})^{-|S|} \\ &= -\frac{1}{|S|} \sum_{i \leq |S|} \mathbb{P}(S_i|S_{<i})\end{aligned}$$

On the other hand, cross-entropy is a way to measure how much the information between two probability distributions differs. It is often used as the loss function for classification tasks in machine learning (Song et al., 2023), including text generation. When one of the distributions is unknown (as is the case when dealing with language modeling), it can be estimated as follows:

$$\mathcal{C}(S) = Loss_M(S) = -\frac{1}{|S|} \sum_{i \leq |S|} \mathbb{P}(S_i|S_{<i})$$

Thus, we can calculate the perplexity for S as the mean cross-entropy for S given a generative model M:

$$\log PP_M(S) = Loss_M(S)$$

Moreover, given that the relation between likelihood and cross-entropy is given by a monotonic function, the relative positions between different data points does not change. This means that we can use the loss from GPT-SW3 (M in this case) to determine the perplexity of a given essay (S in this case). For the rest of this paper we will refer by perplexity to $-\log PP_M(S)$ as opposed to $PP_M(S)$. This is due to the fact that the second number is more likely to underflow as it is a multiplication of probabilities.

[4]An underflow occurs when small numbers are rounded down to zero by the computer due to how floating-point numbers work.

Level	Mean	Median	Std
A1	5.25	5.01	0.78
A2	4.49	4.49	0.74
B1	4.13	4.15	0.48
B2	3.96	3.95	0.42
C1	3.67	3.60	0.36
C2	3.46	3.48	0.48

Table 3: Statistics on the perplexities of GPT-SW3 on the Swell-pilot essays per level. Note that all values get lower the more advanced the students are. This is an indication that as L2 learners advance in their journey, their language approaches that of the language model, which we are assuming should be close to that of an L1 speaker.

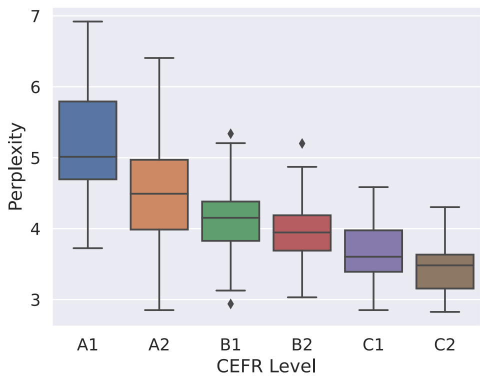

Figure 1: Boxplots for the perplexities of the different CEFR levels. As we can see, as the L2 learner's level progresses, the perplexity of their texts according to GPT-SW3 diminishes.

4 Experiment 1: Perplexity and CEFR Levels

In this section we analyse the perplexities of the essays given by GPT-SW3. We begin by doing statistical analyses of the perplexities by level in Section 4.1. We then do a linguistic analysis of some of the essays of each level in Section 4.2 with the aim of identifying patterns in how the perplexities are distributed within the essay texts.

4.1 Quantitative Analysis

As we can see from Figure 1 and Table 3, the essays from more advanced learners tend to have lower perplexity than those of less advanced learners. This is evidenced when looking both at the mean and the median values of the perplexities for each CEFR level as the more advanced levels have a lower mean and median value.

When looking at the boxplots in Figure 1 we see a similar pattern appear. For each subsequent level, the boundaries of the same quartile are noticeably lower than of the previous level. For example, the first quartile's roof and floor values in level A1 are higher from the ones in level A2, which are in turn higher from the ones in level B1, and so on. The exceptions to this are the boxplots of levels C1 and C2, which have somewhat similar distributions. A possible explanation for this could be that both of these levels are considered to be essentially fluent in the target language, meaning that both kinds of L2 speakers would be able to produce pretty similar sentences. Another possible explanation could be that the normalized essays chosen as a substitute for C2 level essays have a higher perplexity than actual C2 level essays.

However, it is also important to note that the boxplots still have a big overlap between levels, especially in adjacent ones. This means that, while there is a tendency for the perplexities of the essays from GPT-SW3 to get lower the more advanced a learner is, it is by no means a strong indicator for determining the CEFR level of a Swedish L2 learner. This makes sense as language learning itself is a continuous endeavor, as opposed to a discrete one (e.g., Ortega, 2012).

Finally, when looking at the standard deviations, we can see that they also get lower the higher the level. A possible explanation for this could be that the more advanced a learner is, the more likely they are to experiment with the language within certain boundaries that they know to work.

4.2 Qualitative Analysis

To better understand the phenomenon of perplexity, we have also carried out a qualitative analysis. We have selected essays for this qualitative analysis in the following way: we ignored the essays that were outliers in terms of perplexity on each level; of the remaining essays, we picked the two with the highest perplexity, the two with the lowest, and the two closest to the median in their respective levels; and level C2 was ignored from this analysis due to its similarity to the C1 essays. This leaves us with six essays per level for a total of 30 essays.

The analysis has several aspects that we have chosen to focus on. First of all, we want to see what the value of the perplexity depends upon in linguistic terms when seen on a token level and whether this correlates with the CEFR levels. We also want to know whether the perplexities within

Jag heter NN Jag är 16 år Jag bor i Göteborg en dagar är inte bra Därför Jag min kompisa bråkade mid andra alver bråkade fall grap andra alver mycket fel vi spelar fottboll mach hallen fott boll entjejär ringer polisen . Kommer snabb Polisen Polisen är frågar en kille vad händer Kille är inta bra svenska är lite svenska . han förstår inget efter Polisen fråga folk sen är skriver bver Polisen sting hollen fott boll . efter allt bra en dagar ar bra Jag ska gå bibilotek Jag ska fråga en Tjerär Jobbar biliotek Jag vill läsa en bok på lätt svenska hon är hjäpar Jag är läsa svenska bok . Jag träna svenska . myct Svenska lär sig Repetera alltid hemma samma

(a) Original essay in Swedish.

My name is NN I am 16 years old I live in Gothenburg one days is not good Because I my buddyse$ quarreled wihth other tuddents quarreled hole grup other tuddents very wrong we play fottball mach the fott ball hall onegirles call the police . Come soon the Police The Police is ask one guy what happens Guy is nott good Swedish is little Swedish . he understand nothing after the Police ask people then is writes bver the Police thingg the fott ball hall . after all good one days iss good I will go library I will ask one Girrles Work library I want read a book in easy Swedish she is helpes I is read Swedish book . I train Swedish . mucch Swedish learn me Repear always home same day

(b) Translated version of the essay to English.

Figure 2: An example of an A1 level with median perplexity. Darker colors correspond with higher perplexities. Note that the translation was made with the aim of the text being aligned while trying to replicate grammatical errors and misspellings found in the original text.

an essay can be used to help guide or inform on possible aspects on which to focus when grading an essay. Finally, we want to understand how perplexity in LLMs works when dealing with text that was generated by an L2 language learner.

In more practical terms, our intention is to examine whether variations in perplexity can be explained by the linguistic competence of a learner. We focus particularly on sections and tokens with high perplexity, setting a threshold at 6 based on the analysis of graphs in Appendix B, showing how perplexity is distributed across the essays and their variation across different levels. We then analyse the tokens above this threshold across several dimensions.

4.2.1 Placement Within an Essay

The first hypothesis we have explored is that tokens at the beginning of an essay would have higher perplexity values. The idea is that the first few words would be relatively more unexpected for the language model than the text found later in the essay. Looking through the different levels, we can state that this is indeed true in most cases, as can be seen in Appendix A.

This is more noticeable at lower levels, especially if an essay starts with the pronoun *jag*[5] and its various forms. Two examples of this are *Jag heter NN [...]*[6] or **Min skolan ligger [...]*[7]. The lowest perplexity for the first tokens in an essay can be observed in essays starting with the formal subject *Det är / Det finns*[8].

The high perplexity at the beginning of an essay does not seem to characterize essays of a certain level. Therefore, it could be reasonable to ignore the perplexities of the first five or six tokens for successful practical applications of perplexity for L2 essays. Another option would be to weight the perplexity scores depending on their position in the essay.

We also observe that the perplexity values tend, in general, to go down by the end of an essay. This could be because the model knows better what to expect due to the preceding context. Exceptions would arise where other unexpected elements, such as errors, may occur by the end of an essay.

4.2.2 Placement Within a Sentence

The second hypothesis we have explored is that tokens at the beginning of a sentence would have higher perplexity values. It has proven not to be the case.

Essays at levels A1 and A2 can exhibit lack of end-of-sentence punctuation, which makes it next to impossible to separate the increase in perplexity due to the beginning of a sentence with the increase in perplexity due to having a run-on sentence. Essays at levels B1, B2 and C1 do not show regular spikes in perplexity at the beginning of individual sentences. Where such spikes have been observed, this was due to other linguistic reasons, such as errors, rare words, some subjunctions, register switch (from formal to informal or vice versa) or contextually unexpected turn in narration.

Based on this analysis we suggest that the perplexity spikes at the beginning of a sentence could be treated as any other within an essay. This is supported by the fact that GPT-SW3 looks at strings of tokens, which tend to be longer than sentences.

4.2.3 Parts of Speech

Another hypothesis we have explored is that different parts of speech would have different perplexity values in general. The distribution of parts of speech among tokens with higher perplexity shows that content words[9] are much more often perplexing for the model. The percentage of content words of high perplexity is about ∼55-70%. Meanwhile, only ∼15-20% of all the words with high perplexity are function words.[10] The rest of the words with high perplexity are constituted by proper names, numerals, modal verbs, and punctuation.

The high representation of content words suggests a strong impact of semantics, contextual use, and fixed expressions on the probabilities of words expected to be used. A large number of the content words with high perplexity can be explained by various errors, such as non-idiomatic usage, incorrect spelling, or morphological errors. An example of non-idiomatic usage would be *efter allt*, which would be translated word-for-word to English as *after all*. However, this expression is not used in Swedish.

On the other hand, a high perplexity in function words is more often than not triggered by syntax errors, such as missing words or punctuation, issues in word order, and to a lesser degree by misspellings.

An interesting case is presented by high perplexities in multi-word expressions (MWEs). Quite a few of those combine with rare words that appear in combination with just a few other words. Our model is therefore triggered to expect a certain word once the initial part of an MWE is used, such as *å* in the expression *å ena sida..., å andra sidan*.[11] When, the form **å annan sidan* is used instead, the system flags the word *annan*[12] as a perplexing one. In another case, the initial preposition was omitted by a learner from the expression *i alla fall*,[13] so the system flagged *alla* as highly perplexing, whereas the error depended on the omitted token *i*. This same concept can be seen with phrases. That is, perplexity tends to be lower within, say, a noun phrase as words inside it become more predictable.

The last comment on the effects of parts of

[5] This pronoun is the equivalent of the pronoun *I* in English.

[6] Can be translated to English as *My name is NN [...]*.

[7] Can be translated to English as **My the school lies [...]*. Note the use of non-standard language by using a possessive and a determinant on a noun.

[8] Can be translated to English as *There is / There exists*.

[9] Nouns, verbs, adjectives, and adverbs.

[10] Such as prepositions, articles, particles, conjunctions, and pronouns.

[11] This can be translated to English as *on the one hand..., on the other hand*.

[12] The indefinite form of the word *other*.

[13] Can be translated to English as *anyway*.

speech on perplexity needs to be made in connection to proper nouns and names. Names are highly perplexing in general in our data, but even more interesting is the fact that some are significantly more so than others. For example, perplexity for *Kanada* [14] is much higher than for *Afrika* [15]. While we do not have enough proper names in the 30 essays we have selected for qualitative analysis to make any generalisations, we consider that this is a direction that is worth pursuing, especially in relation to possible demographic biases in data and models.

4.2.4 Punctuation

Regarding punctuation, we did not originally expect it to factor significantly into the perplexity of the text. However, we found that the highest spikes have been observed in the use of citation marks and apostrophes. Apostrophes are not used in standard Swedish, which can explain its effect on the model. Meanwhile, the perplexity spikes caused by citation marks could be explained by their low use in the training data for our model. Since citation marks are used at higher proficiency levels (at least in the Swell-pilot data), their high perplexity values may effect the assignment of a CEFR level.

As a take-away lesson, we consider that punctuation in general adds noise and should be exempt from perplexity calculations in connection to essay grading.

4.2.5 Errors

Spikes in perplexities in the running essay text show relatively strong correlation with errors. The majority of words with high perplexity contain some kind of error, either on the token itself (misspelling, morphology, etc) or errors within the previous context (word order, missing punctuation or missing syntactic word, etc).

About ~65-80% of the highly perplexing tokens in essays at A1 and A2 level are related to errors of various types. This number gradually decreases up to the point where at B2 level and higher less than 50% of high perplexity words have a straightforward error associated. In some cases high perplexity may be explained through a (rather vague) notion of non-idiomatic language, use of relatively rare words, deponent verb forms[16], creative compounding, register, abbreviation, etc. The analysis even suggests that word tautology is punished by perplexity, i.e. an overuse of the same content word in close context.

We can summarise this by saying that in the majority of cases, high perplexities reflect an error on the token, or on the previous token. Spelling, morphology, and to a lesser degree syntax are the main reasons of high perplexity in the running text. Wrong word choice, informal register of a word, and non-idiomatic or semantically inappropriate words are also among the errors that can explain higher perplexities in our model.

However, error prediction based on perplexity, is not straightforward, since the high perplexity of a correctly used token may depend on an erroneous usage of the token before. Moreover, errors are not systematically causing high perplexity scores. At lower levels words exhibiting errors with misspellings, capitalisation, morphology and missing punctuation might receive relatively low perplexities. This apparent lack of systematicity could be explained through some of the effects that we have seen in other sections of this analysis, such as lexical choice and frequency effects, the location of the error within the text, among others.

4.2.6 Frequency effects

Given that perplexity is based on probability distributions of the tokens, it is dependent on the frequency of tokens in the dataset on which the language model was trained on.

While we noticed that frequency of vocabulary has a strong correlation with perplexities, a more systematic analysis of word frequencies against perplexity of words in sentence is left for future work.

One of the things that we noticed is that while rare words tend to have higher perplexity values, frequent words like conjunction *och*,[17] the personal pronoun *jag*,[18] and the link verb *att vara*[19] have varying perplexities, depending on their context and neighbouring words.

Another interesting observation with regards to frequency are formulaic expressions that go through language variation. For example, *kommer att* is an expression that can indicate something about the future. A second way to write this would be to drop the *att* particle. However, this second use is not widely spread and is reflected more sparsely

[14] *Canada*
[15] *Africa*
[16] E.g. *bildades*, translated to English as *were built*.
[17] Equivalent to *and* in English.
[18] Equivalent to *I* in English.
[19] Which can be translated to English as the verb *to be*.

Level	Mean	Median	Std
Beginner	4.13	4.09	0.58
Intermediate	4.28	4.32	0.42
Advanced	3.59	3.55	0.45

Table 4: Statistics on the perplexities of GPT-SWE3 on the original Swell-gold essays per level. Even though the beginner-level essays have lower mean and median values when compared to the intermediate-level essays, they have a higher standard deviation.

Level	Mean	Median	Std
Beginner	3.05	3.02	0.28
Intermediate	3.11	3.11	0.26
Advanced	3.10	3.08	0.27

Table 5: Statistics on the perplexities of GPT-SWE3 on the normalised Swell-gold essays per level. Note that all of the statistics from these essays are much more closer to each other across levels when compared to those of the original essays (Table 4).

in online data (Berdicevskis et al., 2024). Our data analysis shows that in cases where *att* has been dropped, the content verb coming after *kommer* gets high perplexity score, as if the model expects *att* but sees a verb instead. In cases where *att* is preserved, the perplexity is on the low level on all tokens.

5 Experiment 2: Perplexity and Text Normalisation

In this section we analyse whether the perplexity of an essay given by GTP-SW3 is reduced when dealing with a normalised version of it. The idea is to establish whether non-standard language correlates with perplexity and to what degree.

When looking at the perplexities in the Swell-gold dataset in Figure 3 and Table 4, we notice that there is not a clear pattern regarding the proficiency level. While this appears to contradict the findings of Section 4.1, this could be due to how the labels were obtained. As mentioned in Section 3.2, these labels were assigned according to the course students are taking, as opposed to actual learner performance.

When comparing the original and the normalised versions of the essays, we see two noticeable tendencies. The first is that, even though the original essays seem to have different distributions depending on their level, the normalised ones have pretty much the same distribution regardless of it, as seen in Tables 4 and 5.

The other tendency is that the perplexity between the original and the normalised versions of the essays go down in all of them. Even though we have suspected this when first looking at Figure 3, the fact that there is an overlap in the boxplots should not be ignored. However, this is confirmed when looking at the figures in Table 6. Here we notice that the minimum difference between the perplexity of the original essays and their normalised versions

Level	Mean	Median	Std	Min
Beginner	1.07	1.03	0.48	0.13
Intermediate	1.16	1.12	0.31	0.52
Advanced	0.49	0.43	0.30	0.05

Table 6: Statistics on the difference between the perplexities of GPT-SWE3 on the original and the normalised Swell-gold essays per level. Note that the minimum values of the difference are all positive, meaning that the perplexity of the normalised essays is always lower than that of their respective original essay.

is still positive, confirming our hypothesis that the perplexity of an essay goes down after its normalisation.

These results corroborate the findings from Section 4.2. That is, the biggest effect of learner language on perplexity comes from errors and the use of non-normative language. This confirms our hypothesis that perplexity is indicative of learner language at different levels.

The remaining spikes in perplexity in normalized essays indicate use of rare words, potentially register switches, citation marks, among others.

6 Conclusions

One of the issues with large language models tends to be their lack of interpretability and explainability. This keeps true with generative models such as those based on the GPT architecture despite them being able to generate text "justifying" their reasoning (Blevins et al., 2023).

In this work we aimed to explore the relationship between the perplexity from a decoder-only model of Swedish and the complexity of the text of an L2 speaker's essays.

We found that there is an inverse relationship between the CEFR level of an essay and its perplexity. However, due to the overlap between the values of each level means that they are not a strong indica-

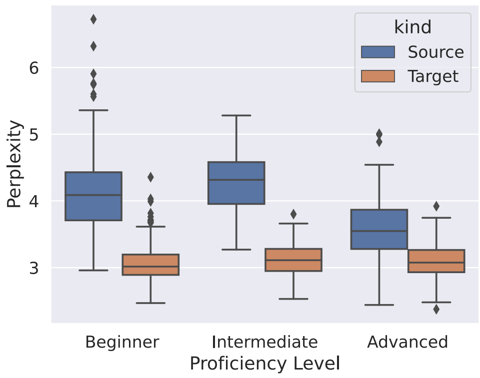

Figure 3: Boxplots for the perplexities of the different proficiency levels. Even though there does not seem to be an obvious pattern between the levels of the texts and their perplexities, the normalised texts show a much lower perplexity than the original texts. Moreover, the distributions of the normalised texts are much more similar to each other than to their original versions.

tor for the level of the essay. Moreover, we found evidence that proficiency levels derived from the course a student is taking might not be indicative of the actual proficiency of the essays.

We also found that there are perplexity effects through the essays that are not exclusive to L2 language, such as placement within a text, punctuation, frequency of the tokens, among others. Despite that, some of the more prevalent effects are characteristic of L2 language, such as errors, non-idiomatic use of the language, and multi-word expressions.

There is a correlation between the use of non-standard language as an L2 language learner. This conclusion can be drawn by the fact that the perplexity for every essay became lower after being normalised.

One of the possible applications of these could be done through the use of these features to help guide human graders with the assessment of learner language. The idea of this is to take either a human-in-the-loop (Wu et al., 2022) or a hybrid intelligence approach to evaluation (Dellermann et al., 2019). However, it would be of essence to disentangle the perplexity effects that are specific to L2 speakers from those effects that are not. This would allow us to have a more reliable and fair estimation. This, however, remains to be explored in the future.

CEFR are categorical classes used to describe language proficiency for teaching and assessment convenience. Despite that, language development itself works as a continuum, where essays within each particular level are not homogeneous with regards to their linguistic complexity. This continuum of linguistic complexity of learner language has rather vague and arbitrary cut-off points between one level and another (Hulstijn et al., 2010; Ortega, 2012; Alfter et al., 2021). Given the context of our experiments, we hypothesise that the perplexity score per essay can help place each essay on a scale between one level and another and that it may be an indirect way of grading essays within the same level. However, this is a hypothesis that needs to be explored in another paper.

7 Limitations

Throughout this paper we have talked about perplexity as a way to measure the surprisal of a model. While this is a useful way to interpret this value in an intuitive manner, it is important to note that this is just a metaphor. We are not treating the language model as an agend and humanising it. This is particularly relevant as they still have a vast amount of limitations and their misuse can lead to undesirable results (Weidinger et al., 2022).

8 Ethical Considerations

In high stakes situations such as those related to language learning it is important to constantly audit our systems and processes to ensure that unfairness does not begin to creep into the process. Moreover, we consider that a human-in-the-loop approach is the correct way to go about, as mentioned in Section 6. This allows the students to ask both for explanations on the results and for a revision of these in case they consider them to be erroneous.

Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure Nationella språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

This work is part of a continuous endeavour on exploring the use of machine learning techniques to study second language learner texts in Swedish. Moreover, we aim to modernise the L2 tools that are currently available through the research infrastructure of Språkbanken Text, in particular via Språkbanken's learning platform Lärka,[20] which

[20] https://spraakbanken.gu.se/larka/

allows researchers, as well as teachers and learners, to interact and analyse these kinds of texts in an automated manner.

The second author is also supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the *Centre for Linguistic Theory and Studies in Probability (CLASP)* at the University of Gothenburg.

References

David Alfter, Therese Lindström Tiedemann, and Elena Volodina. 2021. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. *The Northern European Journal of Language Technology (NEJLT)*, Vol.7 No.1:1–35.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Aleksandrs Berdicevskis, Evie Coussé, Alexander Koplenig, and Yvonne Adesam. 2024. To drop or not to drop? predicting the omission of the infinitival marker in a swedish future construction. *Corpus Linguistics and Linguistic Theory*, 20(1):219–261.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprint*, arXiv:2005.14165.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

COE Council of Europe. 2001. *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643.

Simon Dobnik, Mehdi Ghanimifard, and John Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 1–11, New Orleans. Association for Computational Linguistics.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2023. GPT-SW3: An autoregressive language model for the nordic languages. *arXiv preprint*, arXiv:2305.12987.

Simon Flachs, Ophélie Lacroix, and Anders Søgaard. 2019. Noisy channel for low resource grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 191–196, Florence, Italy. Association for Computational Linguistics.

Jan H Hulstijn, J Charles Alderson, and Rob Schoonen. 2010. Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, pages 11–20.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Julia Klezl, Yousuf Ali Mohammed, and Elena Volodina. 2022. Exploring linguistic acceptability in Swedish learners' language. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 84–94, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Murathan Kurfalı and Robert Östling. 2023. A distantly supervised grammatical error detection/correction system for swedish. *Swedish Language Technology Conference and NLP4CALL*, pages 35–39.

Alice Kwak, Cheonkam Jeong, Gaetano Forte, Derek Bambauer, Clayton Morrison, and Mihai Surdeanu. 2023. Information extraction from legal wills: How well does GPT-4 do? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4336–4353, Singapore. Association for Computational Linguistics.

Shalom Lappin. 2023. Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information*, 33(1):9–20.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Jingcheng Niu and Gerald Penn. 2020. Grammaticality and language modelling. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 110–119, Online. Association for Computational Linguistics.

Official Journal. 2016. Recital 71. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, L 119.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 technical report. *arXiv preprint*, arXiv:2303.08774.

Lourdes Ortega. 2012. Interlanguage complexity. *Linguistic complexity: Second language acquisition, indigenization, contact*, 13:127.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.

Ming Qian. 2023. Performance evaluation on human-machine teaming augmented machine translation enabled by GPT-4. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 20–31, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners*. White Paper. Open AI.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2023. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint*, arXiv:2302.13971.

Elena Volodina. 2024. On two SweLL learner corpora – SweLL-pilot and SweLL-gold. *Huminfra Conference*, pages 83–94.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).

Elena Volodina, Ildikó Pilán, and David Alfter. 2016b. Classification of Swedish learner essays by CEFR levels. In *CALL communities and culture – short papers from EUROCALL 2016*, pages 456–461. Research-publishing.net.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Trans. Intell. Syst. Technol.*, 12(5).

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The nordic pile: A 1.2tb nordic dataset for language modeling. *arXiv preprint*, arXiv:2303.17183.

A Perplexity Plots for the Beginning of the Essays

In this appendix we present typical 'perplexity shapes' for the beginning of a sentence. In Figure 4 we present the plots for the first 25 tokens of the essays from Figure 5 with the exception of the one at level C1.

Figure 4: Perplexity plots for the first 25 tokens of the sample essays from Figure 5. The X-axis shows the running number of a token, while the Y-axis shows the perplexity score. Relative perplexity for the first several tokens is stably high, with a few exceptions. Essays at C1 and C2 level exhibit the same tendency.

B Perplexity Plots of the Essays

In Figure 5 we present plots of the perplexity changes throughout some of the essays. These plots were used to help inform a cut-off line between what we consider relatively high and relatively low perplexity values.

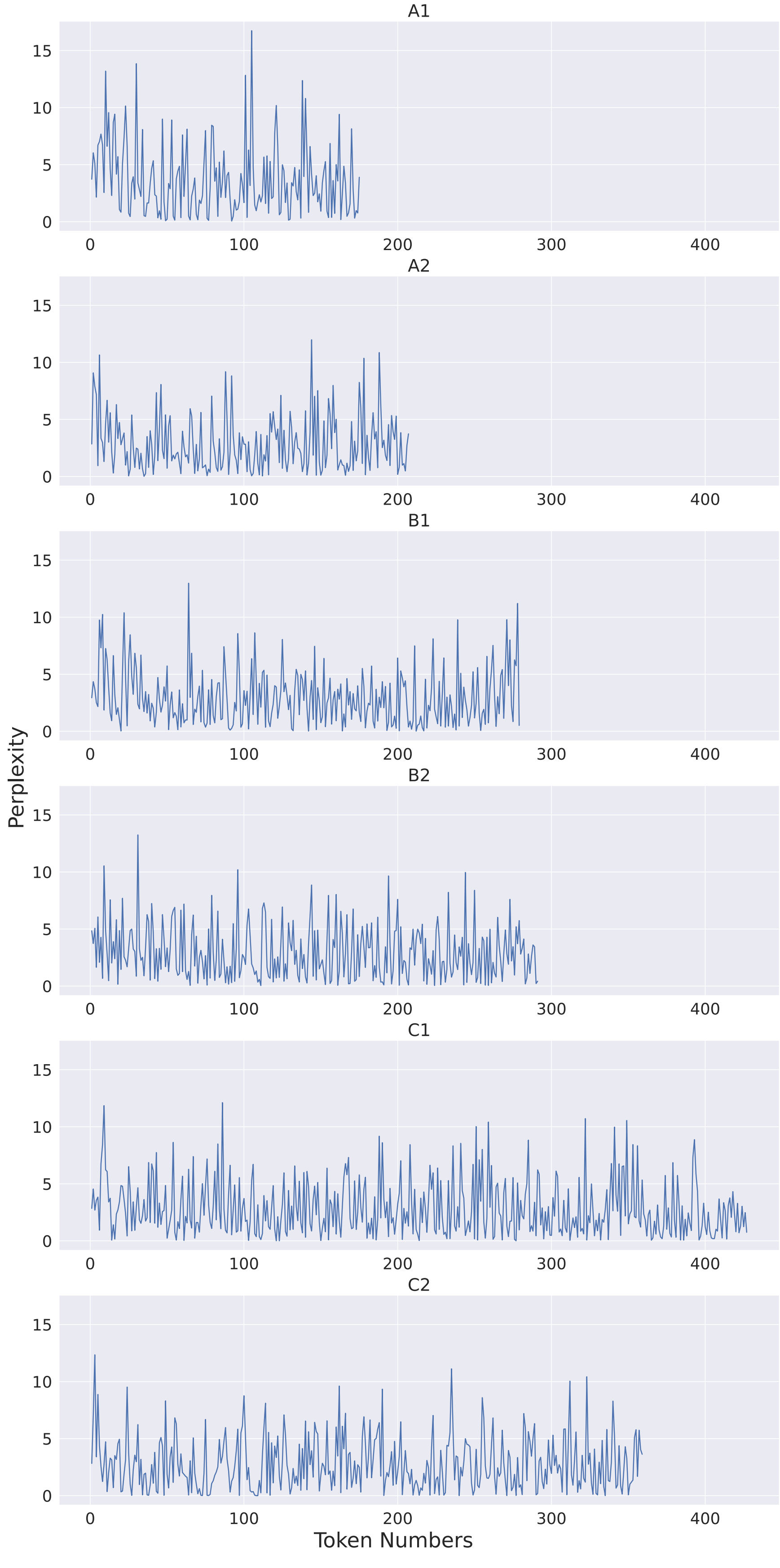

Figure 5: Sampled perplexity shapes for full essays with median perplexity at levels A1 to C1. The X-axis shows the running number of a token, while the Y-axis shows the perplexity score.