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

Ricardo Muñoz Sánchez[1], Simon Dobnik[2], Elena Volodina[1]

[1]Språkbanken Text, University of Gothenburg, Sweden [2]CLASP, FLoV, University of Gothenburg, Sweden

UNIVERSITY OF GOTHENBURG

The Idea

Are LLMs sensitive to linguistic characteristics of L2 learner texts?

- We use perplexity as a measure of surprisal
- Our hypothesis is that perplexity is related to the complexity of the language in L2 learners' essays in terms of level
- We also analyze the relation between perplexity and linguistic features of L2 learner language

The Data

We used two different datasets from the Swell corpora collection. They consist of second-language learner essays of Swedish which have been processed and annotated in different ways.

Swell-Pilot [Experiment 1]

Consists of 502 essays collected between 2012 and 2016. They have been anonymized and annotated for CEFR level.

Swell-Gold [Experiment 2]

Consists of 502 essays collected between 2017 and 2021. They have been pseudonymized and annotated with the course level the students were in (beginner, intermediate, or advanced). It contains both the original essays and normalized versions of them.

The Model

We use GPT-SW3, an auto-regresive model based on the GPT architecture. It was trained on The Nordic pile, a 1.3TB internet corpus in the Nordic languages. It is the best-performing generative model in Swedish as of now.

The Metric

Perplexity is how likely it is that an observation of a sample is made by an estimator. Thus, we can intuitively interpret the perplexity as a way to measure how "surprised" the model is to see a given sequence.

Perplexity

Given that $\mathbb{P}(S)$ can be seen as the probability of a token given its previous context, we have that:

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

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

Cross-Entropy

Cross-entropy is a way to measure how much the information between two probability distributions differs and is often used as the loss function for classification tasks in machine learning. When one of the distributions is unknown, it can be estimated as follows:

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

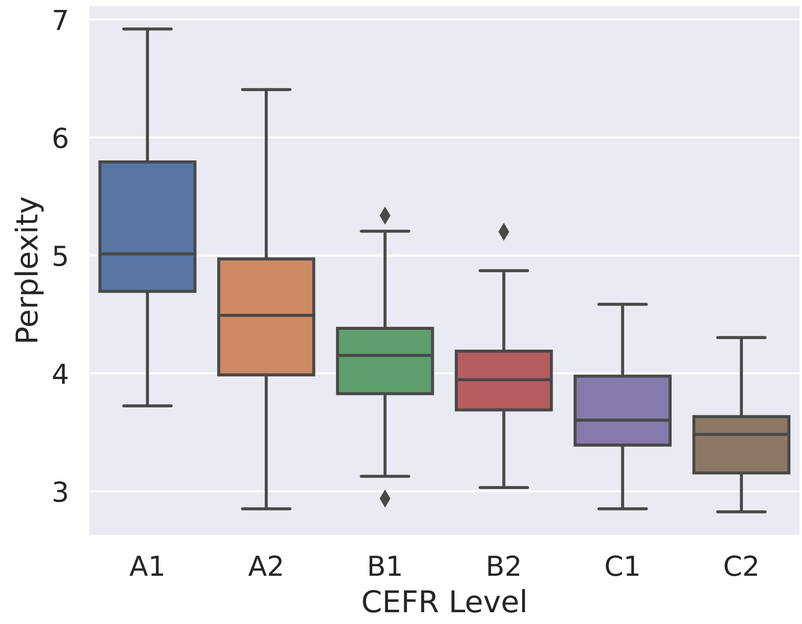
Perplexity and Cross-Entropy

Perplexity tends to be a very small number, so we risk having an underflow in our calculations. Given that cross-entropy is the logarithmic version of perplexity, we can use this instead. The relation is monotonic, so the relative positions between different datapoints does not change.

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

Perplexity and CEFR Levels [Experiment 1]

As the CEFR level of the essays progresses, the perplexity of their texts according to GPT-SW3 diminishes. Note that, even though there is a downwards tendency, the boxplots still overlap.

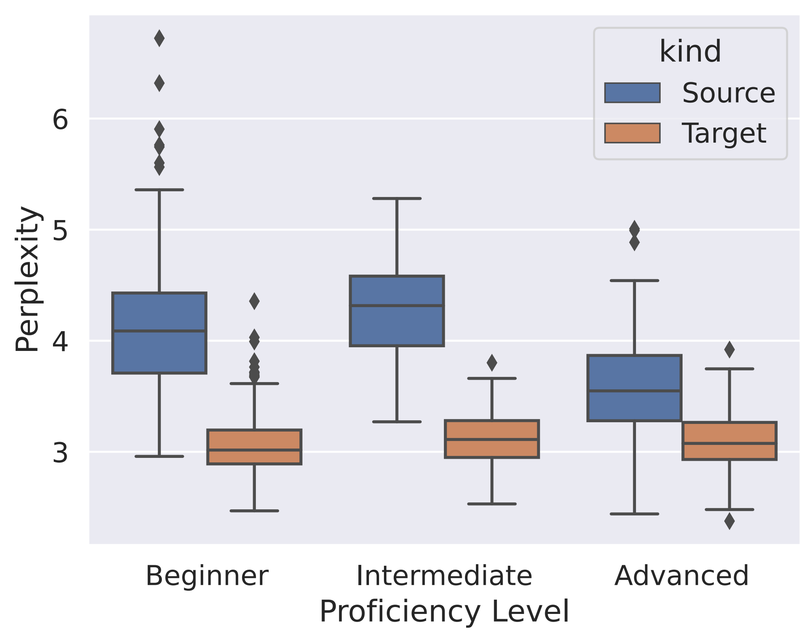


Causes for High Perplexity [Experiment 1]

- **Placement within an essay** - the earlier a token appears, the higher its perplexity regardless of its level.
- **Placement within a sentence** - placement within sentences has negligible effect on the perplexity of a token due to the context window.
- **Parts of speech** - high perplexity in content words tends to be related to non-idiomatic usage, while function words tend to always have high perplexity.
- **Punctuation** - apostrophes and quotation marks show very high perplexity.
- **Errors** - errors within an essay usually lead to perplexity spikes and are strongly related to essay level.
- **Frequency** - rare and very common words lead to higher perplexity.

Perplexity and Normalization [Experiment 2]

Normalizing (or standardizing) the language in the essays reduces their perplexity, indicating that L2 learner language tends to increase perplexity. Moreover, normalization tends to make texts more similar to each other in terms of perplexity.



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. 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.