UNIVERSITY OF GOTHENBURG

SPRÅKBANKENTEXT

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

Ricardo Muñoz Sánchez, Simon Dobnik, Elena Volodina


Introduction

The Idea

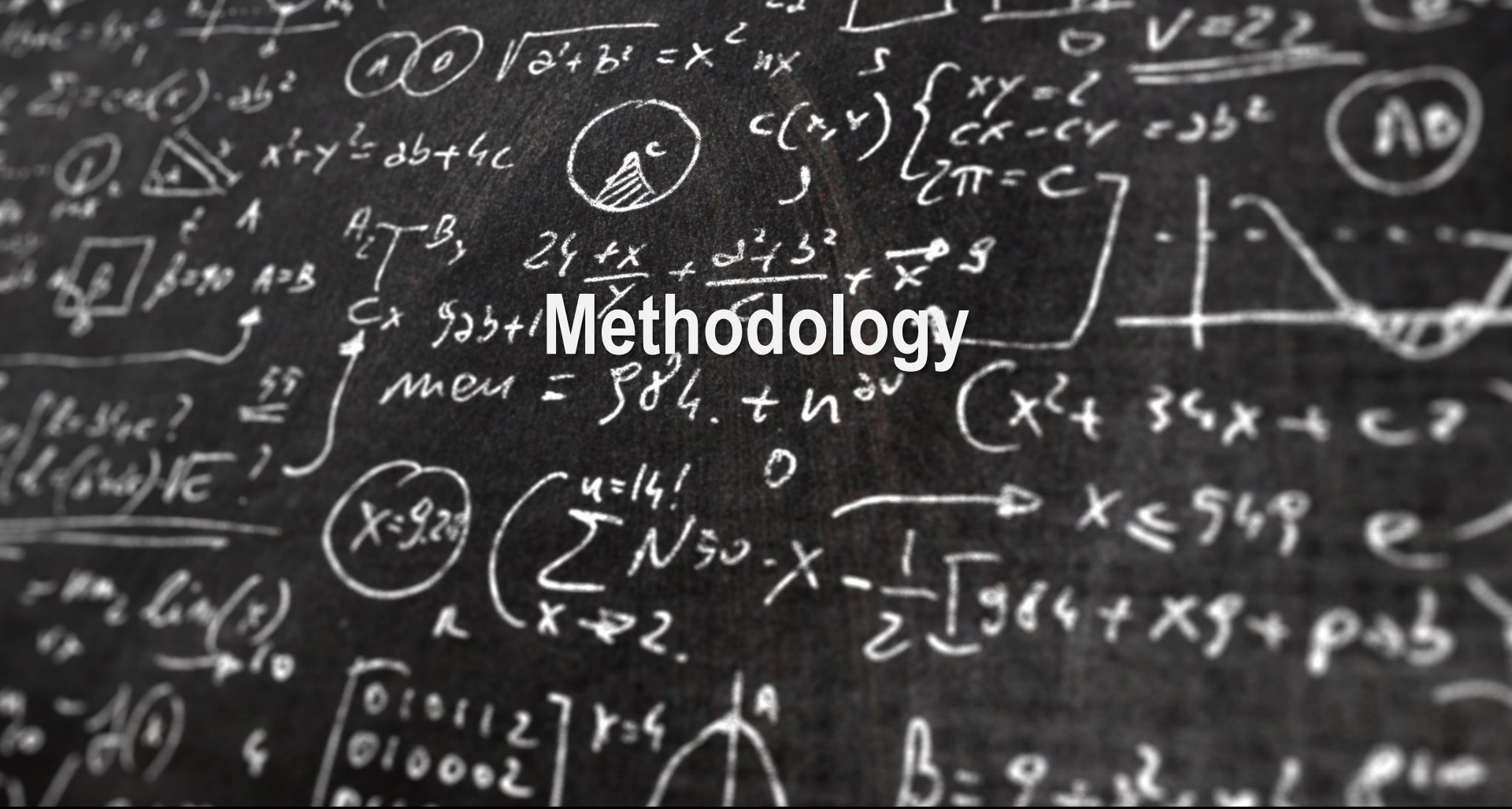
Neural models often perform better than feature-based approaches

However, they are obscure and not easy to interpret

This is an issue in high-stakes situations, such as second language assessment

Our Approach

- Perplexity has been used as a measure of surprisal
- We want to see how it interacts with L2 learner texts of Swedish
- Our hypothesis was that perplexity would be related to the complexity of the language in L2 learners' essays



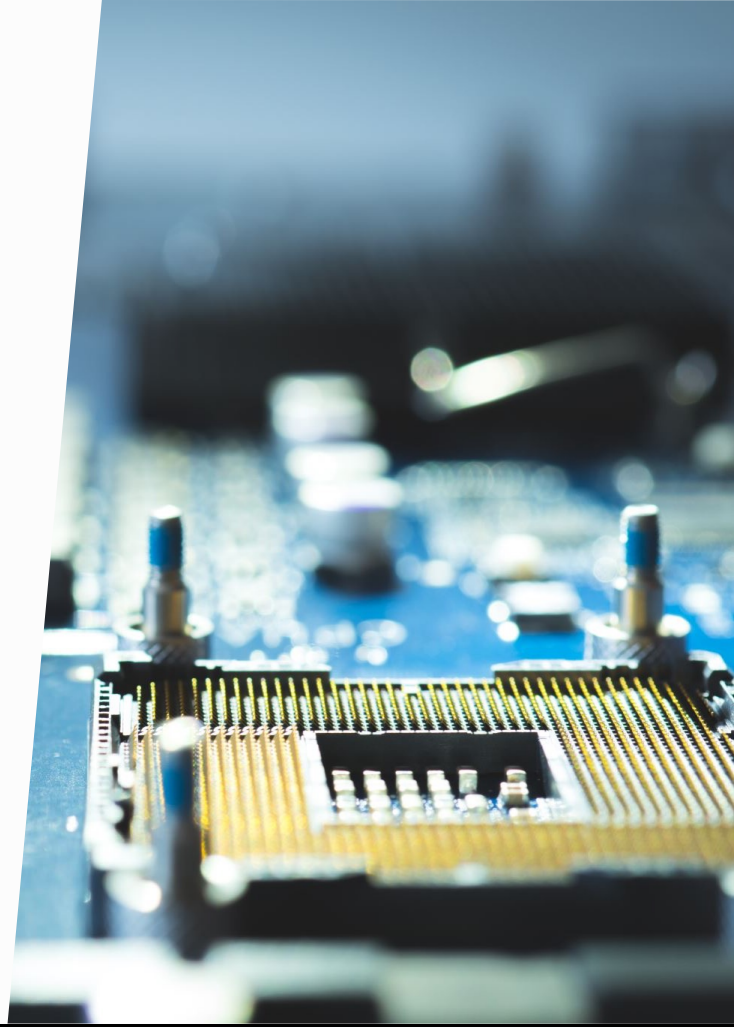
Methodology

Our Methodology

- Take student essays from the Swell-Corpora
- Feed these essays to GPT-SW3
- Analyse how perplexity is distributed within the essays

GPT-SW3

- Auto-regressive model based on the GPT architecture
- Trained on The Nordic Pile, a 1.3TB internet corpus in the Nordic languages
- Is currently the largest and best-performing generative model available for Swedish



Swell

- Swell-Pilot
  - 502 essays
  - Collected between 2012 and 2016
  - Annotated for CEFR level

- Swell-Gold
  - 502 essays
  - Collected between 2017 and 2021
  - Contains original and normalized versions of the essays
  - Annotated with course level of the students

Perplexity as a Measure of Surprisal

- Perplexity is the probability that an estimator sees an observation
- When dealing with generative models, it is the probability of a token given the previous context
- We can also use cross-entropy as it is in logarithmic scale and less likely to underflow

Perplexity as a Measure of Surprisal

- Definitions
  - S is a sentence
  - S_i is the i-th token of the sentence
  - $S_{<i}$ is the sentence up to the i-th token, but not including it

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

- Cross-Entropy:
  - $\mathcal{C}(S) = \log PP(S) = -\frac{1}{|S|} \sum_{i \leq |S|} \log \mathbb{P}(S_i | S_{<i})$


Experiment 1
Perplexity and CEFR Levels

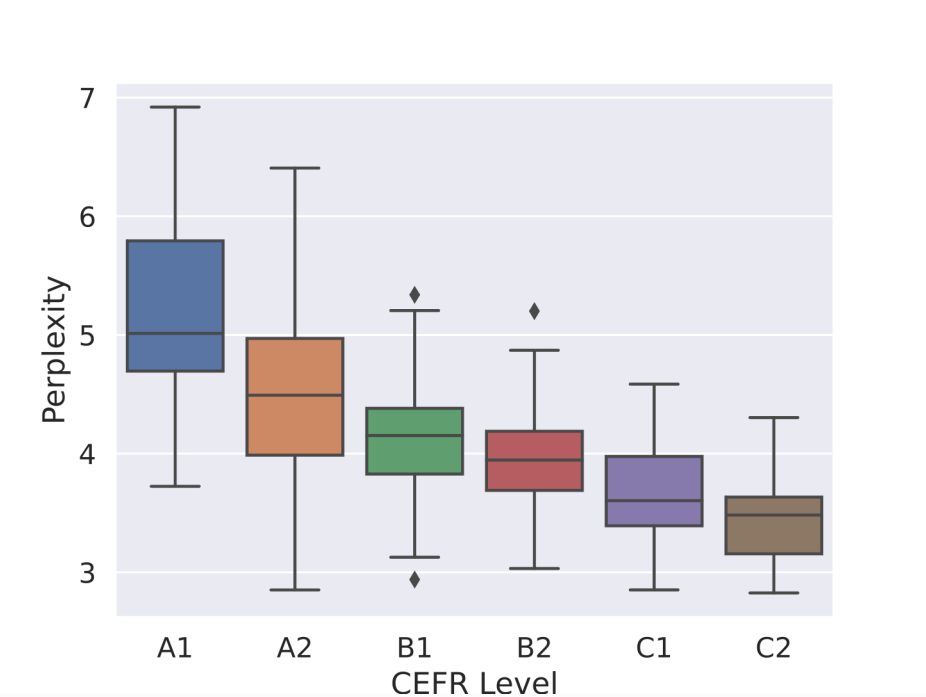
Comparing Perplexity Across CEFR Levels

We take essays annotated for CEFR levels and calculate their perplexity at essay and token levels

We compare essay-level perplexity for the different essays

For each essay we check the perplexity for each token and compare it with different hypotheses we made beforehand

Perplexity and CEFR Level



Qualitative Analysis

- Placement within an essay
  - Tokens at the beginning of essays tend to have higher perplexity regardless of level

- Placement within a sentence
  - The location of a token within a sentence does not affect perplexity
  - The major exception to this are run-on sentences

Qualitative Analysis

- Parts-of Speech
  - Content words with high perplexity tend to be related to non-idiomatic usage
  - Function words tend to have higher perplexity regardless of usage

- Punctuation
  - Citation marks and apostrophes tend to have high perplexity
  - Apostrophes are not used in standard Swedish

Qualitative Analysis

- Errors
  - There appears to be a strong correlation between errors and perplexity spikes
  - This is also related to essay level

- Frequency Effects
  - Perplexity is based on (conditional) probability
  - Rare words tend to have high perplexity
  - Very common words also have high perplexity


Experiment 2
Perplexity and Normalization

Comparing Original and Normalized Essays

We compare the perplexities of the original versions of essays with their normalized versions

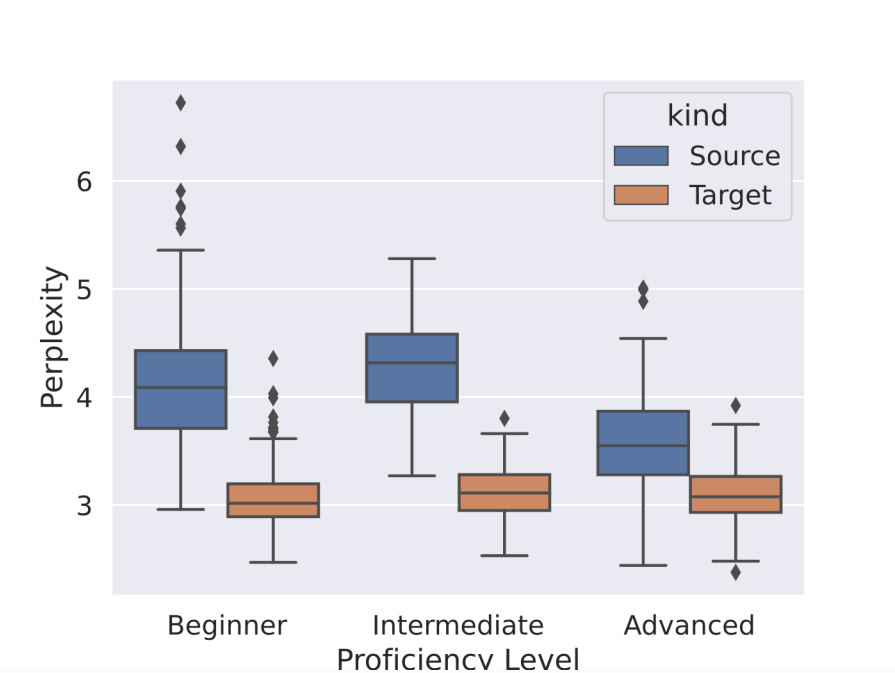
We check what is the effect of normalizing the text of the essays on their perplexity

We check whether course level is a good proxy for CEFR levels

Perplexity and Normalization

- The normalized versions of the essays all have similar perplexities
- The perplexity of a normalized essay is always lower than that of the original one
- There is no clear pattern according to the course levels

Perplexity and Normalization




Conclusions

Conclusions

- There is an inverse relationship between CEFR levels and perplexity
- Course level is not a good proxy for proficiency of the essays
- High perplexity is not exclusive to L2 language
- Non-standard use of language by L2 learners seems to be correlated with higher perplexity

GÖTEBORGS
UNIVERSITET

SPRÅKBANKENTEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@svenska.gu.se
rimusa.github.io