Project outline: Advanced analysis and computational modeling of reading data

Project overview:

For my project I will use the GECO-corpus to implement two models that predict the gaze duration of words recorded while reading English text. Gaze duration is defined as the sum of all fixations on the current word during the first-pass reading, before the eye moves out of the word. The goal is to analyze the influence of syntactic versus semantic features on the ability to predict gaze duration. To do this I will compare two models.

Model A: Linear prediction model using syntactic features

This model will predict the gaze duration of words based on rather simple syntactic features. Those features will be:

- Word length
- Position in the sentence
- Word frequency

The gaze duration will be modeled as a weighted sum of these features. A grid search will be used to optimize the feature weights. However, if the search space is too large or inefficient, I will switch to gradient descent to find optimal weights more efficiently. This model serves as an interpretable and transparent baseline, helping to isolate the predictive value of low-level lexical properties.

Model B: ANN using GPT-embeddings

This model will be an ANN using GPT-embeddings as input to predict gaze duration of words. The embeddings for every word will be extracted using the

OpenAI API. These are designed to capture deep semantic meaning and context, not just simple syntax.

The goal is to train this ANN to predict gaze duration as accurately as possible based on semantic features.

Goal

The comparison between the two models will highlight whether semantic features (Model B) contribute significantly to predicting gaze duration beyond what can be explained by syntactic features alone (Model A).

Hypothesis

This project is guided by the following hypotheses:

- H1: Semantic embeddings significantly improve gaze duration prediction.
- H2: Syntactic features already explain a substantial portion of gaze duration.

Optional Model C: Fusion of Syntax and Semantics

To further strengthen the analysis, a third model may optionally be implemented. This model combines the syntactic features from Model A (e.g., word length, position, frequency) with the semantic embeddings from Model B. These combined features will be fed into a single artificial neural network.

The purpose of this hybrid model is to explore whether the combination of both information types leads to a higher predictive performance than either model alone.