Project notes

Project Goal

- Analyze *gaze duration* during reading
- Compare **syntactic** vs. **semantic** features in prediction
- Dataset: **GECO corpus** (English)

What is Gaze Duration?

- The total time the eyes fixate on a word during first-pass reading, before moving to another word

  ➔ Used to understand cognitive processing during reading, related to mental effort

Model A – Linear prediction model using syntactic features

- Input features:
  • Word length
  • Position in sentence
  • Word frequency
- Output: Gaze duration (as weighted sum)
- Optimization via **grid search or gradient descent if search space gets too large**
- Purpose: Transparent and explainable **baseline model**

Model B – ANN with GPT Embeddings

- Input: GPT-based **word embeddings** via OpenAI API
- Captures **semantic meaning and context**
- Purpose: Predict gaze duration based on semantic features

Goal of Comparison

- Determine if **semantic features** improve prediction over syntax alone

- Assess how much added value semantic information provides

Hypotheses

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

Optional Model C – Hybrid Approach

- Combines syntactic (A) and semantic (B) inputs
- Unified ANN using both feature types
- Purpose: Test whether combining both leads to the best prediction performance