Analysis of Disordered Speech using the APROCSA Dataset

In this assignment, you will analyze connected speech data from individuals with post-stroke aphasia using the APROCSA dataset. The dataset contains speech transcripts and expert consensus ratings on a range of auditory–perceptual features describing different aspects of disordered speech.

Dataset Access: `https://langneurosci.org/aprocsa-dataset/`

You will work with:

- Transcripts of connected speech (in CHAT format).
- Consensus perceptual ratings on a 0–4 severity scale for multiple speech characteristics.

Part A: Descriptive Analysis of Transcripts

This part focuses on understanding the dataset, preprocessing the transcripts, and computing simple descriptive statistics that characterize disordered speech.

A1. Dataset understanding (6 marks)

Briefly describe the APROCSA dataset (2 marks per question):

a) What type of speech data does it contain?

b) What do the perceptual ratings represent?

c) What does the 0–4 severity scale indicate?

A2. Transcript preprocessing (10 marks)

1. Extract only the participant's speech from each transcript (e.g., the lines beginning with `*PAR:`).
2. Clearly define:
   - what you consider an **utterance**,
   - what you consider a **token** (word),
   - how you treat **filled pauses** (e.g., *uh*, *um*),
   - how you treat explicit **pause markers** such as `(.)` or `(...)`.
3. Provide one short **before and after example** showing how a raw transcript line is transformed after preprocessing.

A3. Descriptive statistics (10 marks)

For each participant and investigator, compute the following measures:

- Total number of utterances
- Total number of tokens (words)
- Mean utterance length (in tokens)
- Lexical diversity (e.g., Type–Token Ratio or another clearly defined metric)
- Number of filled pauses per 100 tokens
- Number of pause markers per 100 tokens
- Present your results in a single table where rows correspond to participants/investigators and columns correspond to these measures.
- Write a short note on how the speech of investigators and participants differ.

A4. Linking transcript measures to perceptual ratings (10 marks)

Choose any **three** APROCSA perceptual rating dimensions that you believe should be reflected in transcript statistics (for example: *Pauses within utterances*, *Short and simplified utterances*, *Anomia*, *Reduced speech rate*).

For each chosen rating:

1. Identify one transcript-derived measure as a proxy.
2. State a hypothesis about the expected relationship.
3. Compute a correlation (Pearson or Spearman).
4. Interpret the result in 2–3 sentences.

Summarize your results.

A5. Visualization (10 marks)

Produce **one** of the following:

- a heatmap of APROCSA ratings (participants $\times$ features),
- scatter plots showing your three correlations,
- or a bar plot of selected transcript measures across participants.

Include a short caption explaining what the figure shows.

Part B: Regression Model for Predicting a Perceptual Rating

This part introduces predictive modeling and interpretation of feature effects.

B1. Target variable (2 marks)

Choose one APROCSA perceptual rating to predict (e.g., *Pauses within utterances*, *Anomia*, or *Short and simplified utterances*). Briefly justify your choice in 2–3 sentences.

B2. Feature design (10 marks)

Construct at least **six** transcript-derived features, including:

- Number of utterances
- Number of tokens
- Mean utterance length
- Lexical diversity
- Pause markers per 100 tokens
- Filled pauses per 100 tokens

You may add additional features if you wish to do so.

B3. Regression model (10 marks)

Fit a regression model to predict the chosen perceptual rating using your transcript-derived features. Use **leave-one-out cross-validation (LOOCV)**.

Finally report the following:

- the true and predicted ratings for each participant,
- an error metric such as MAE or RMSE,
- optionally, the correlation between predicted and true values.

B4. Feature interpretation (10 marks)

Standardize your features and report the regression coefficients. Explain:

- which features increase predicted severity,
- which features decrease predicted severity,
- which features appear most influential.

Write a short paragraph interpreting at least three features.