

# Computational Psycholinguistics Assignment 1

**Nidhi Vaidya**

**2023114005**

## A1. Dataset understanding

- a) What type of speech data does it contain?
  - Dataset contains speech transcripts, not isolated words/sentences
  - Speech produced by post-stroke patients with aphasia
  - Transcript stored in chat format (Contains speaker turns, pauses, disfluencies, annotations)
- b) What do the perceptual ratings represent?
  - Ratings are expert auditory-perpetual judgements
  - Capture speech characteristics (pauses, anomia, fluency, simplicity)
  - Reflects how severe a symptom sounds to trained listeners
- c) What does the 0–4 severity scale indicate?
  - 0 = no impairment
  - 4 = severe impairment
  - Higher value = greater impairment severity
  - 5 options on the scale, therefore the scale is ordinal and not continuous

## A2. Transcript preprocessing

### 2. Definitions

- Utterance: one \*PAR line
- Token (word): Words separated by whitespaces, lowercase, punctuation not included
- Filled pauses (uh, um): Special tokens
- Pause markers: Removed from text, counted separately as pause events

### 3. Example:

\*PAR: I (.) I was uh going to the store (...)

After preprocessing: i i was uh going to the store

Changes:

The \*PAR tag was removed. One \*PAR line considered as one utterance, everything converted to lowercase.

Every word is considered as one token and punctuation is ignored.

“Uh” is not ignored but considered as a special token.

Pause markers have been removed.

## A3. Descriptive statistics

File	Speaker	Utterances	Tokens	Mean Utterance Length	TTR	Filled pauses per 100 tokens	Pause Markers per 100 tokens	Filled Pauses Count	Pause Markers Count	Types
aproc sa1554 a.cha	PAR	204	1390	6.81	0.365	11.22	4.89	156	68	507
aproc sa1554 a.cha	INV	151	895	5.93	0.399	2.12	0.11	19	1	357
aproc sa1713 a.cha	PAR	186	1654	8.89	0.387	7.80	1.69	129	28	640
aproc sa1713 a.cha	INV	119	867	7.29	0.488	1.27	0.00	11	0	423
aproc sa1731 a.cha	PAR	398	2652	6.66	0.401	3.32	1.55	88	41	1063
aproc sa1731 a.cha	INV	253	1200	4.74	0.567	0.5	0.25	6	3	680
aproc sa1738 a.cha	PAR	173	1421	8.21	0.427	8.23	2.25	117	32	606
aproc sa1738 a.cha	INV	122	890	7.30	0.466	2.7	0.00	24	0	415
aproc sa1833 a.cha	PAR	211	1792	8.49	0.382	11.33	3.74	203	67	685
aproc sa1833 a.cha	INV	213	1495	7.02	0.456	1.54	0.07	23	1	682
aproc sa194 4a.cha	PAR	314	2734	8.71	0.290	3.66	1.46	100	40	792
aproc sa194 4a.cha	INV	82	594	7.24	0.512	1.18	0.34	7	2	304

Summary of the results:

1. Number of utterances for the participants are always greater than or equal to those of the investigators. This is because since the participants are patients suffering from aphasia, they require more sentences/turns to convey information. Factors like anomia/breakdowns in speech also contribute to the same.

2. Number of tokens for PAR are more than INV, which is expected since number of utterances are more.

- Observation: for case 1833a, INV has more utterances but PAR has more tokens (exception). This could indicate that the investigator spoke more often but said less per turn/utterance. What I could understand from this is: More

investigator turns might indicate greater communicative support, which is consistent with more severe discourse difficulty.

3. PAR a higher mean utterance length than INV. This could be because participants with aphasia often produce long, effortful turns containing multiple pauses and fillers, which remain within a single utterance boundary. But INV frequently contribute short prompts and backchannels distributed across many turns, resulting in shorter mean utterance lengths.

4. TTR (type-token ratio) is always  $PAR < INV$ . This is because INV has a richer and varied vocabulary. Meanwhile PAR (aphasiacs) have lower vocabulary, repeat common words, and anomia may lead to avoidance of specific lexical terms.

5. Filled pauses are way greater for PAR than INV. Due to aphasiacs' disordered speech, anomia, delays.

6.  $PAR \ggg INV$  for pause markers too. Same reasons: PAR breakdown in fluency, difficulty sequencing speech. INV pause deliberately and minimally.

7. For filled pauses,  $PAR \ggg INV$ . Absolute number of fillers more for PAR. INV fillers are rare and incidental.

8. Pause markers are also way greater for PAR. Aphasic speech involves frequent silent/marked pauses. Same reasons as stated above. INV don't pause much, unless waiting for PAR to speak. Also, I noted that some of the INV have 0 pauses. Again supports the same explanation.

9.  $PAR > INV$  for types. Even if PAR repeats words a lot, sheer volume increases the chance of new types. More tokens  $\rightarrow$  more word types (true mechanically).

#### A4. Linking transcript measures to perceptual ratings

Chosen perceptual rating dimensions:

1. Pauses within utterances
2. Reduced speech rate
3. Anomia

1. Identify one transcript-derived measure as a proxy

Perceptual Rating Dimension	Transcript derived proxy measure
Pauses within utterances	Pause markers per 100 tokens
Reduced speech rate	Filled pauses per 100 tokens
Anomia	Type Token Ratio

2. State a hypothesis about the expected relationship.

### Hypothesis 1: Pauses

*Participants with higher ratings for "Pauses within utterances" will exhibit higher rates of pause markers per 100 tokens.*

Expected: positive correlation

### Hypothesis 2: Speech Rate

*Participants with higher ratings for "Reduced Speech Rate" will exhibit higher rates of filled pauses per 100 tokens.*

Expected: positive correlation

### Hypothesis 3: Anomia

*Participants with higher ratings for "Anomia" will exhibit lower rates of type token ratio (TTR).*

Expected: negative correlation

### 3. Compute a correlation (Pearson or Spearman).

Choice: Spearman coefficient since the perceptual ratings are ordinal and not continuous.

Spearman correlations (PAR only):

Pauses within utterances vs pause markers per 100 tokens:  $\rho = -0.655$ ,  $p = 0.158$

Reduced speech rate vs filled pauses per 100 tokens:  $\rho = -0.169$ ,  $p = 0.749$

Anomia vs lexical diversity (TTR):  $\rho = -0.463$ ,  $p = 0.355$

### 4. Interpret the result in 2–3 sentences

Small sample size ( $n = 6$ ) -> low statistical power, non-significant p-values expected

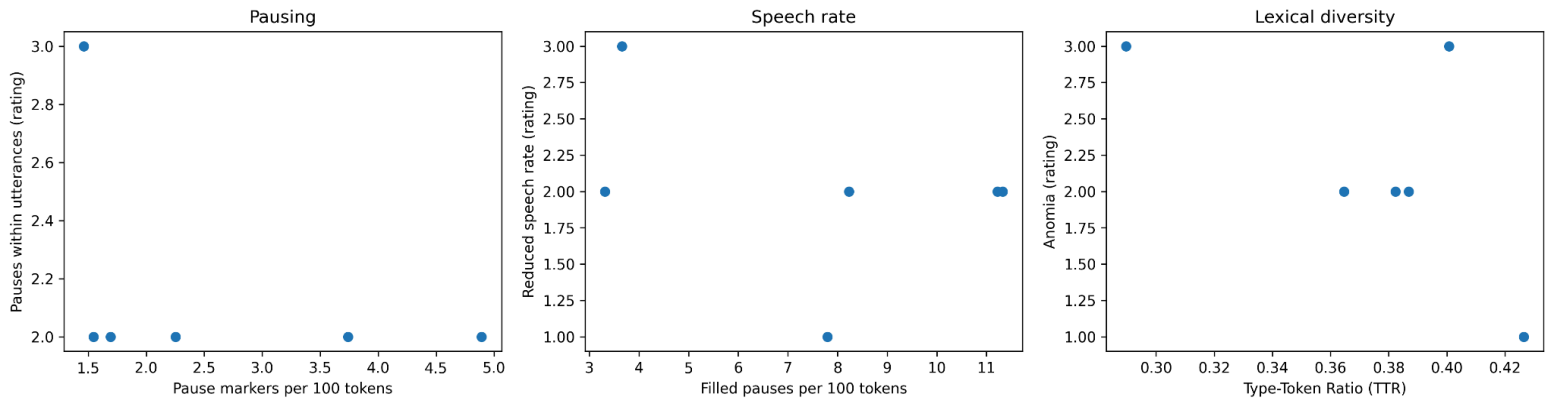
Pause markers vs perceptual pausing: moderate negative trend -> perceptual pausing reflects overall disruption, not just explicit pause annotation

Filled pauses vs reduced speech rate: very weak relationship -> speech rate influenced by silent pauses, planning, articulation speed

TTR vs anomia: negative trend -> higher anomia associated with lower lexical diversity

Overall: transcript measures capture partial but incomplete aspects of perceptual severity

### A5: Visualization



**Figure (A5).** Scatter plots showing the relationship between transcript-derived measures (x-axis) and perceptual ratings (y-axis) for PAR speech across participants: pause markers per 100 tokens vs *pauses within utterances* (left), filled pauses per 100 tokens vs *reduced speech rate* (middle), and type-token ratio (TTR) vs *anomia* (right). Each point represents one participant; overall patterns are weak/moderate with substantial variability due to the small sample size.

#### B1: Target variable

My target variable is **Anomia**

- it already showed a meaningful trend in A4
- links well to TTR, pauses, utterance length
- easy to interpret in B4

#### B2. Feature design

Feature	Why it's relevant to Anomia
Number of utterances	reflects fragmentation of speech
Number of tokens	overall speech quantity
Mean utterance length	shorter turns -> word-finding difficulty

Lexical diversity (TTR)	direct proxy for lexical access
Pause markers / 100 tokens	hesitation due to retrieval problems
Filled pauses / 100 tokens	classic symptom of anomia

- these features capture both quantity and quality of speech
- are theoretically linked to lexical retrieval difficulty

### B3. Regression model

I trained a Ridge Regression model using LOOCV (each participant held out once), and reported predictions on the held-out participant.

### **Results:**

<b>Participant</b>	<b>True Anomia</b>	<b>Predicted Anomia</b>
1	2	1.89
2	2	1.79
3	3	2.49
4	1	1.82
5	2	1.95

6	3	3.25
---	---	------

### Error metric:

Mean Absolute Error (MAE): 0.325

Correlation (Spearman):  $\rho = 0.802$ ,  $p = 0.055$

### B4. Feature interpretation

#### Results:

utterances +0.148

tokens +0.147

mean\_utterance\_length -0.048

ttr -0.131

pause\_markers\_per\_100\_tokens -0.010

filled\_pauses\_per\_100\_tokens -0.071

Features that increase predicted severity (positive):

utterances (+0.148) -> more fragmented turns -> higher predicted anomia

tokens (+0.147) -> slight increase (small effect)

Features that decrease predicted severity (negative):

ttr (-0.131) -> higher lexical diversity -> lower predicted anomia (most meaningful negative)

filled pauses /100 (-0.071) -> small negative here (likely overlap with other predictors)

mean utterance length (-0.048) -> longer utterances -> lower predicted severity

pause markers /100 (-0.010) -> near-zero effect

Most influential (by |coefficient|):

utterances (~0.148) and tokens (~0.147) (largest positives)

ttr (~0.131) (largest negative + most interpretable clinically)

others are minor (especially pause markers)

#### Interpretation:

Using standardized Ridge coefficients, utterances (+0.148) and tokens (+0.147) are the strongest positive predictors, so more fragmented/more speech slightly increases predicted anomia severity. TTR (-0.131) is the most meaningful negative predictor, indicating higher lexical diversity corresponds to lower predicted anomia. Filled pauses (-0.071) and mean utterance length (-0.048) have smaller negative effects, while pause markers (-0.010) contribute almost nothing. Because the dataset is very

small ( $n=6$ ) and features overlap, these coefficients should be interpreted cautiously as relative tendencies rather than definitive causal effects.

Mainly:  $n = 6$  + correlated features  $\rightarrow$  sensitive coefficients