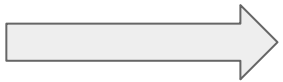
Pattern of Children's Language Production

Delores Tang

Research Question

- Do children across **ages (5-11)** produce languages with different structures (e.g., in terms of **POS tags**)?
- Does the trend differ between children with **typical** and **impaired** language abilities?

Input	Output
• Words (Embeddings) • POS tags (Vectorized) • POS tag frequencies	• Age • Language Impairment



Background

Nouns Before Verbs

- Infants tend to learn nouns first and then verbs
- Waxman et al..2014
- Verbs are critical for language learning
- Language impaired - the effect of mental disorders
  - Spelling, Vocabulary
  - Inefficient use of sentences and utterances
  - Impaired abilities to understand languages

Background

Two Hypothesis

- **Natural Partition:**

  pattern emerged because of the abstract/concrete distinction in how humans perceive objects and events.

  PERCEPTION

- **Linguistic Relativity**:

  English language is Noun-focused.

  ENGLISH

Background

Two Hypothesis

- **Natural Partition:**

  pattern emerged because of the abstract/concrete distinction in how humans perceive objects and events.

  PERCEPTION

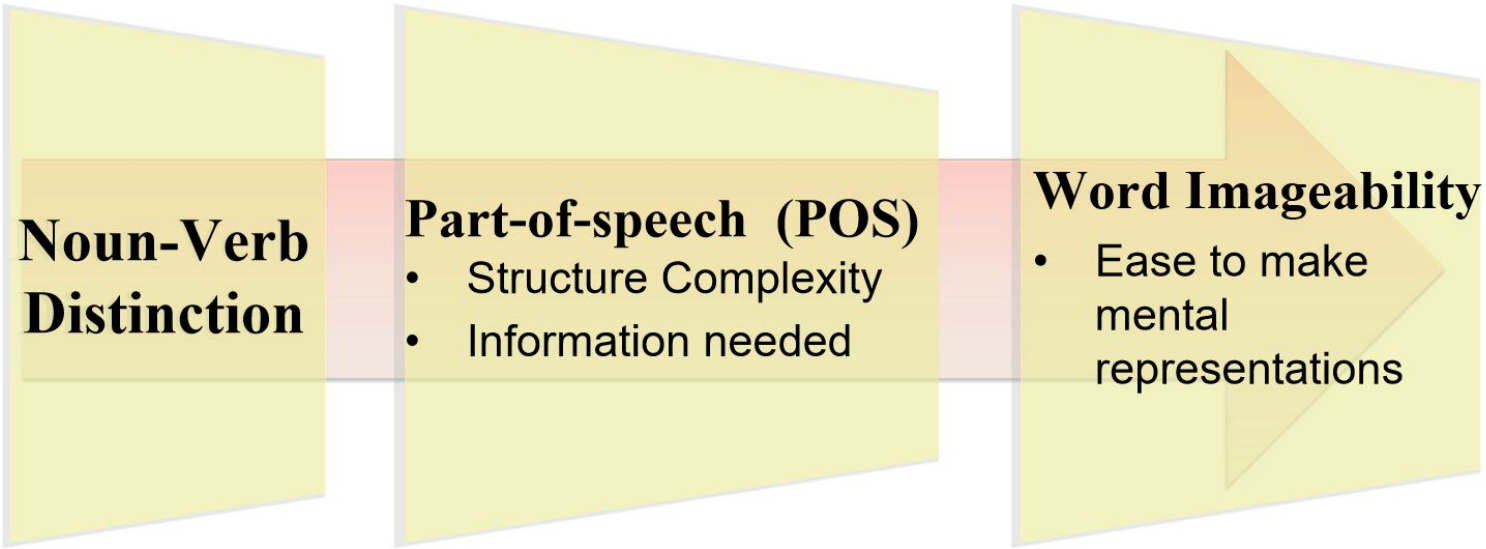
- **Linguistic Relativity**:

  English language is Noun-focused.

  ENGLISH

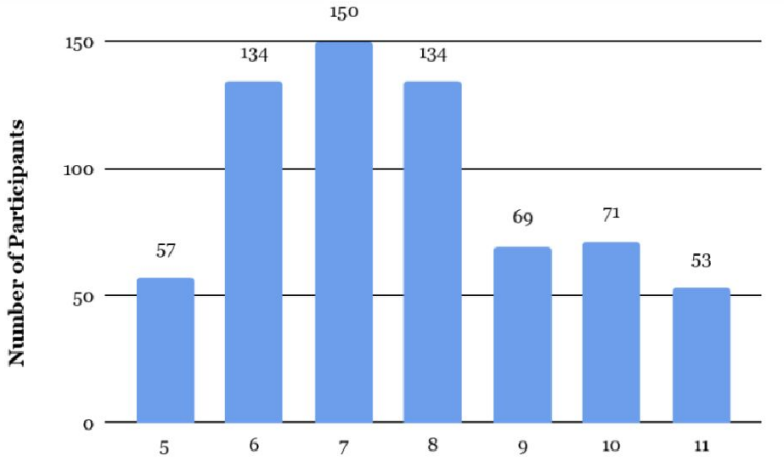
Background

Trend in Research



Data

- **Childes** - a database with raw data from hundreds of studies.
- **Gillam Corpus**
  - Originally used for building Test of Narrative Language (TNL)
  - McDonald's storytelling, natural environment
  - Aged 5-11
  - **Typical**:171; **Impaired**: 497



Data

ID	Gloss	POS tags	Impaired	Age	POS tags frequencies	N:V	$\frac{N+V}{Total}$
0	“I love bacon”	[pro:sub, v, n]	0 or 1	5	……. (42 unique tags)		

Methods

Preliminary Analysis

Output: Age

- ❏ **Noun-Verb Ratio**
  One-way ANOVA
- ❏ **Noun+Verb/Total**
  Least squares
- ❏ **All POS/Total**
  Least squares

Impair vs. Typical
t-test

Word-Based Analysis

Word Embeddings:

- ❏ **GloVe**
- ❏ Word2Vec

Output: Impairment

- ❏ Naive Sequential
- ❏ Simple RNN
- ❏ **RNN with LSTM**
- ❏ Random Forest

Output: Age

- ❏ Simple RNN
- ❏ Random Forest

POS-Based Analysis

Vectorizer:

- ❏ TF-IDF vectorizer

Output: Impairment

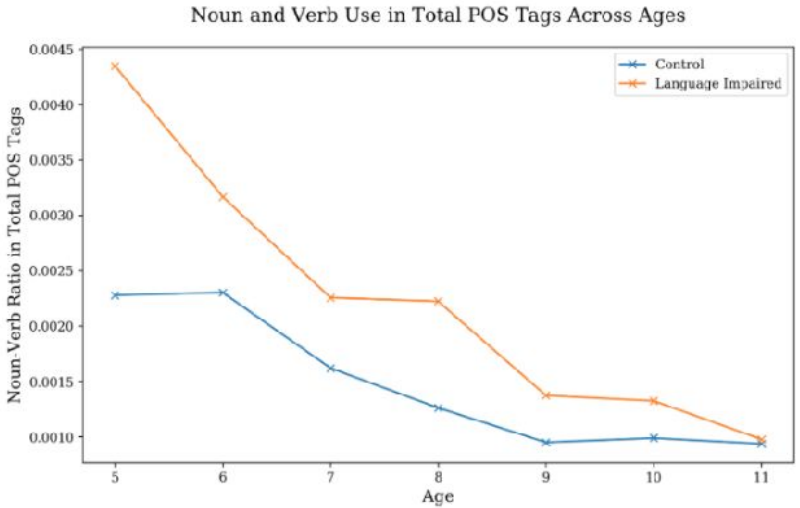
- ○ Logistic Regression
- ○ Naive-Bayes
- ○ Linear SVM
- ○ Ridge/Lasso/Elasticnet
- ○ Random Forest

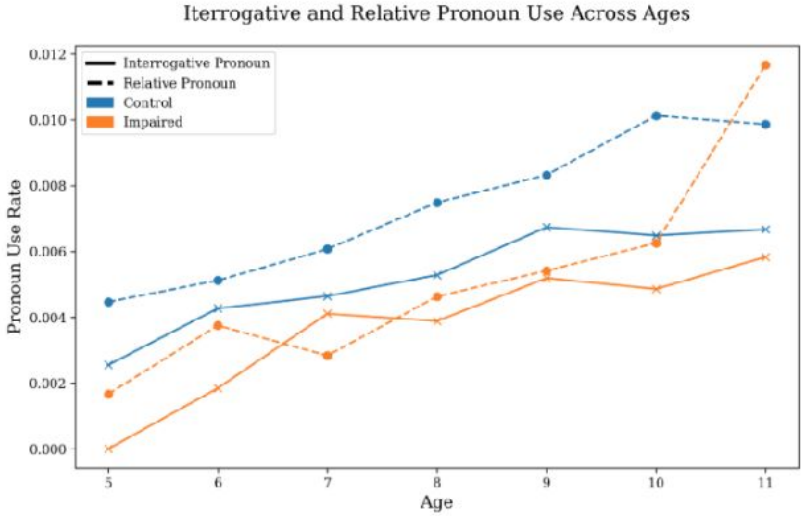
Output: Age

- ○ Logistic Regression

Preliminary Analysis

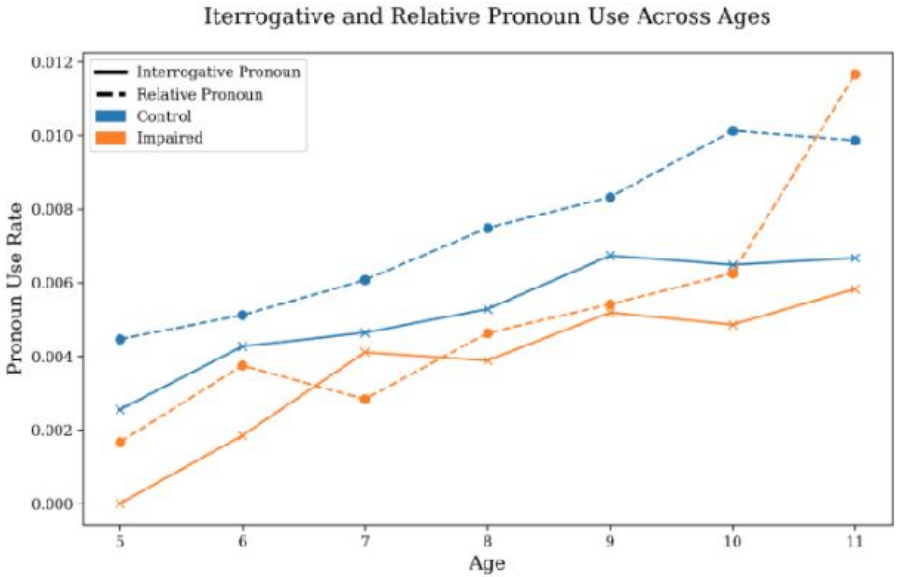
- **N:V ratio** insignificant in relation to **Age**
- **N+V** was negatively correlated to **Age**
- **Interrogative** ('what') and **relative** ('where') **pronouns** are more frequently used as **Age** increases





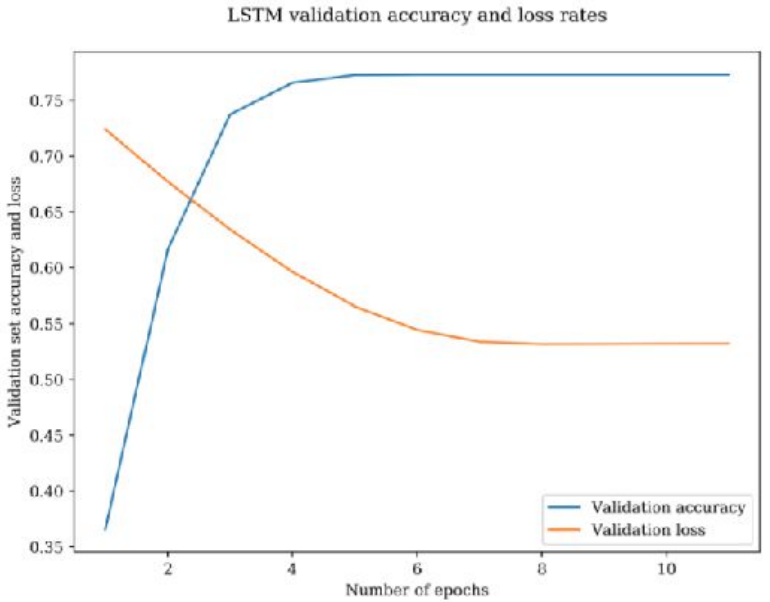
Preliminary Analysis

- Children use more complex sentence structures (more clauses) as they grow older.
- This trend differs for children with typical and impaired language abilities.
- The differences seem to converge as both groups grow older.



Word-Based Analysis

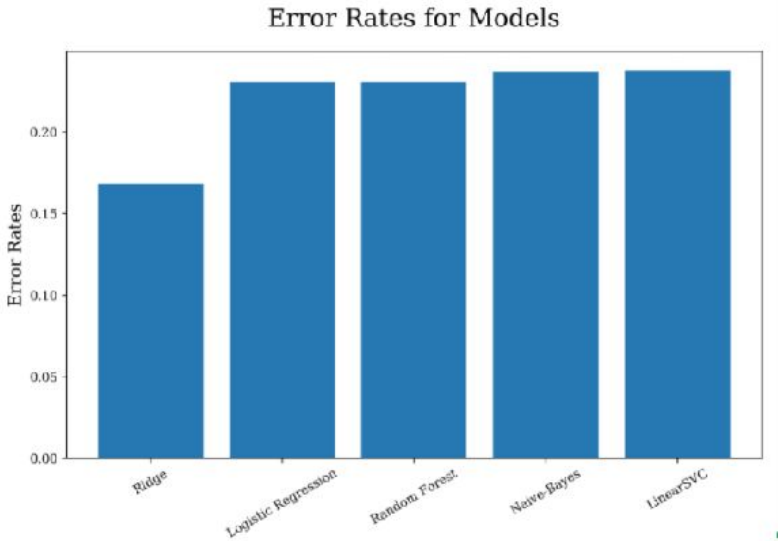
- **GloVe** embedding returns better results.
- **Simple RNN stacked with LSTM** as the best model



- However, validation accuracy stopped increasing at 0.7731 due to insufficient data

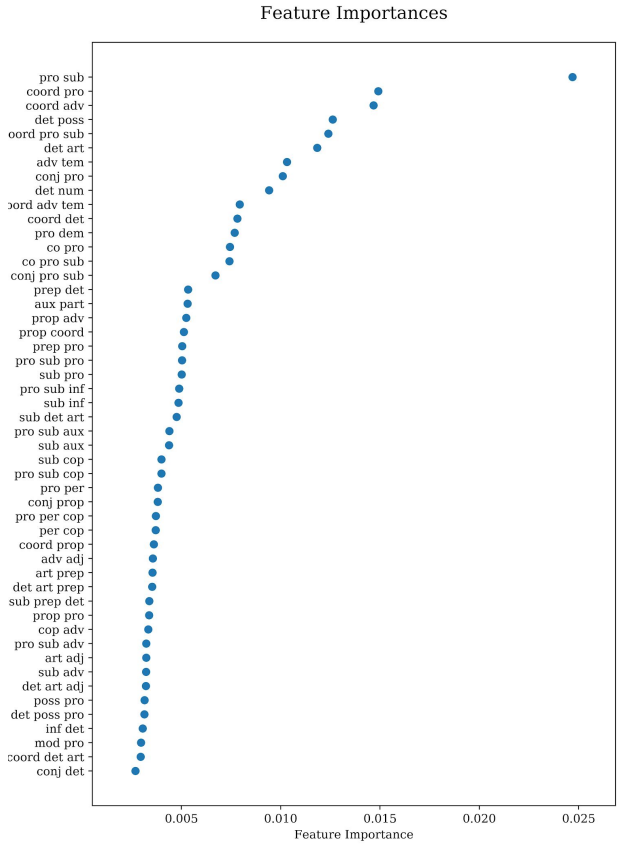
POS-Based Analysis

- **TF-IDF Vectorizer** - ngrams = (1,3)
- Model selection - **Ridge**



POS-Based Analysis

- **Feature importance:**
  - Subjective pronoun
  - Conjunction, pronoun
  - Conjunction, adverb
- Specific types of clauses are learned before others
- More linguistic analysis is needed.



Thanks you very much!