Part-of-Speech Tags, Age, and Language Impairment - A Language Production Analysis using Machine Learning Techniques

University of Chicago
Delores Tang

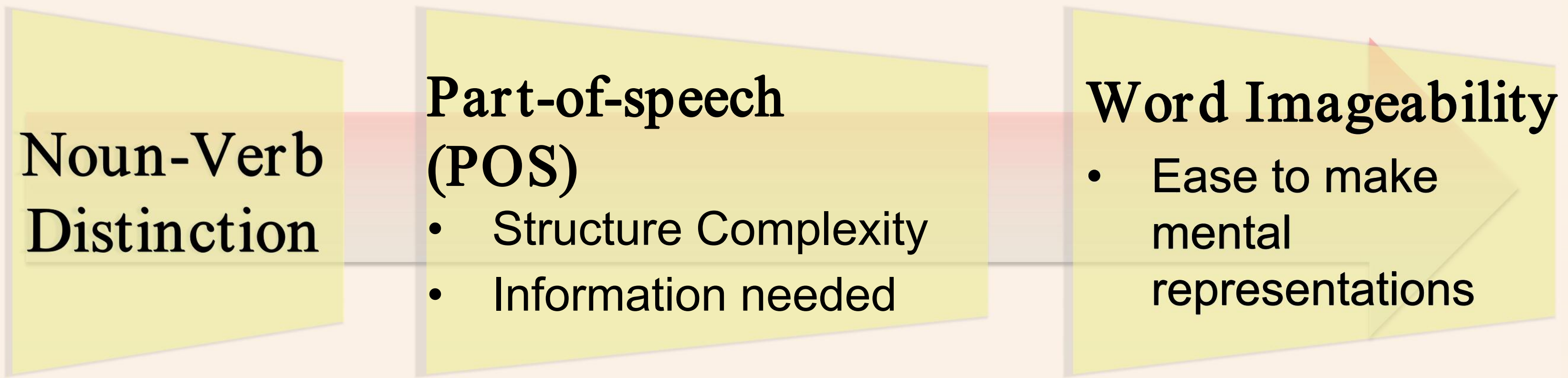
Introduction

"Nouns before Verbs"

- ◆Children's ability to acquire **verbs** seemed to lag behind their ability to acquire **nouns**
- ◆Past literature primarily focused on children under age of 5.

Natural Partition vs. Linguistic Relativity

- ◆**Natural Partition**: pattern emerged because of the abstract/concrete distinction in how humans perceive objects and events.
- ◆**Linguistic Relativity**: English language is Noun-focused.
  - ➢ Was not supported by cross-cultural studies.



Noun-Verb Distinction

Part-of-speech (POS)
- Structure Complexity
- Information needed

Word Imageability
- Ease to make mental representations

Language Impairment

- ◆Diagnosis for **Specific Language Impairment** (SLI):
  - ➢ Receptive: information processing, understanding;
  - ➢ Expressive: spelling & vocabulary, simpler sentences; inappropriate/insufficient utterances.
- ◆Lack of evidence on specific linguistic structures.

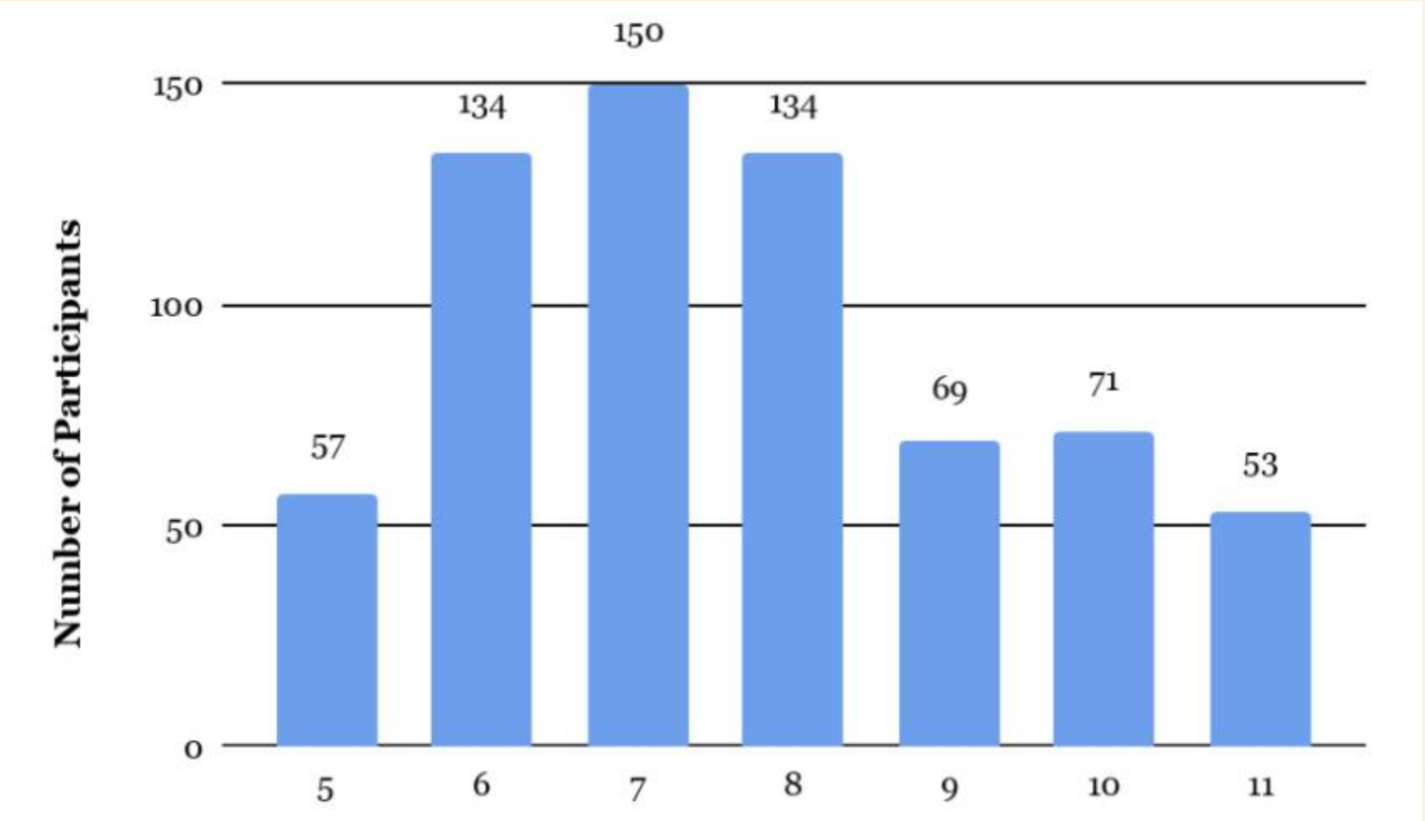
Research Question:

- ◆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?

Methods

Dataset

- ◆**Childes** (Child Language Data Exchange System)
- ◆**Gillam Corpus**
  - ➢ Test of Narrative Language (TNL)
  - ➢ McDonald's storytelling
  - ➢ **Typical**:171; **Impaired**: 497



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)		

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
- LSTM
- Random Forest

Output: Age
- Simple RNN
- Random Forest

POS-Based Analysis

Vectorizer:
- TF-IDF vectorizer

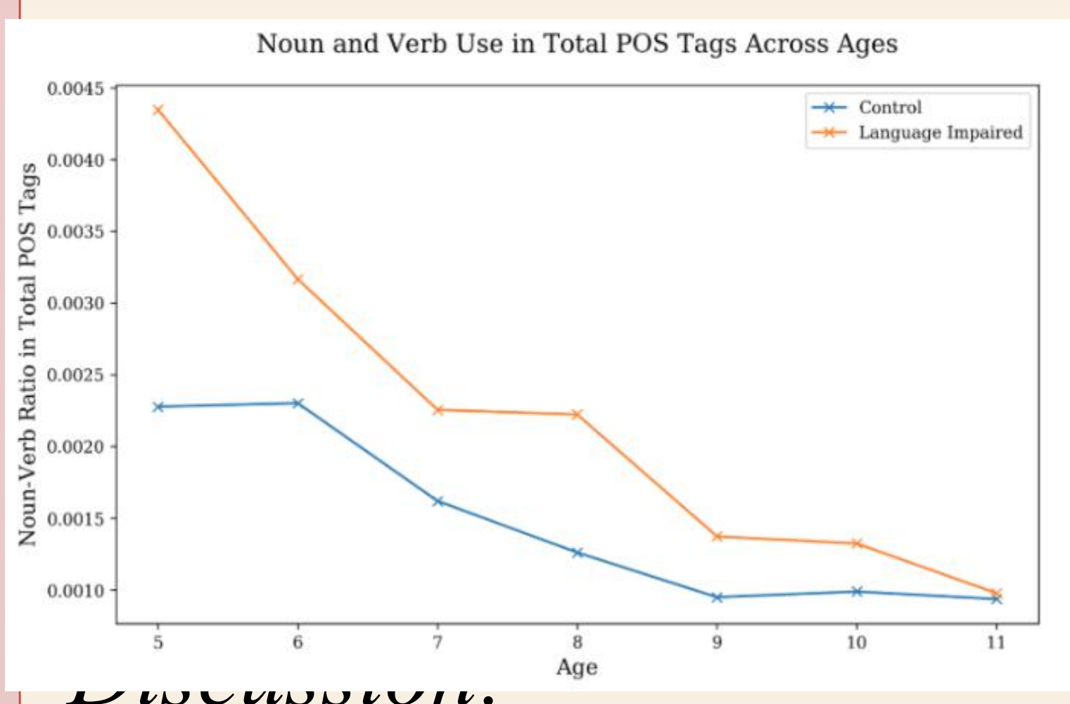
Output: Impairment
- Logistic Regression
- Naive-Bayes
- Linear SVM
- Ridge/Lasso/Elasticnet
- Random Forest

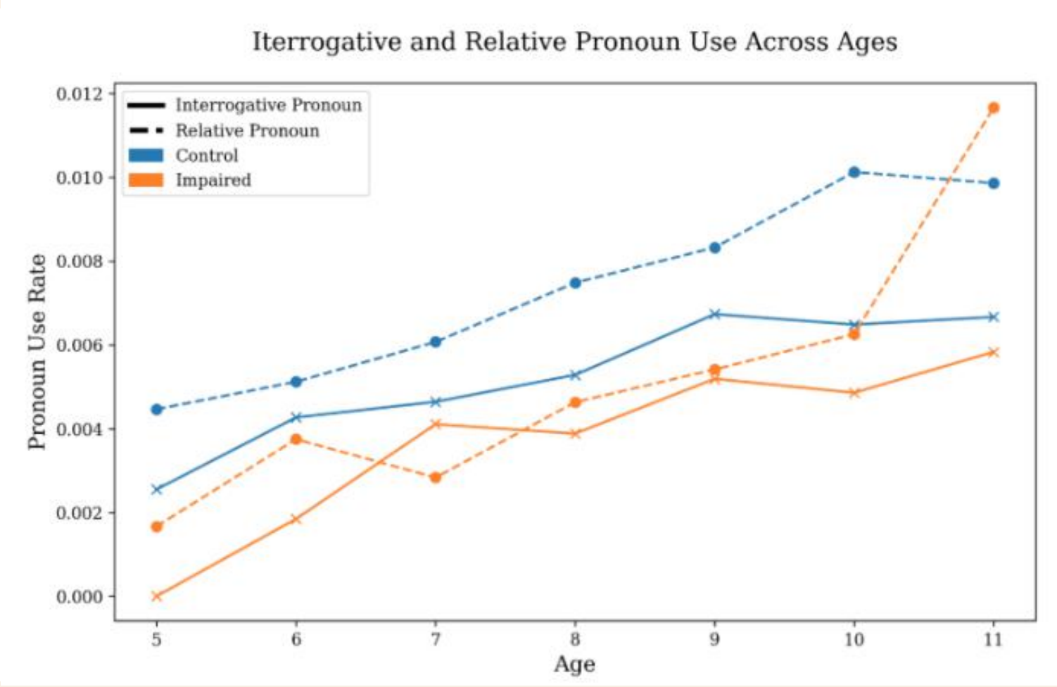
Output: Age
- Logistic Regression

Results & Discussion

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



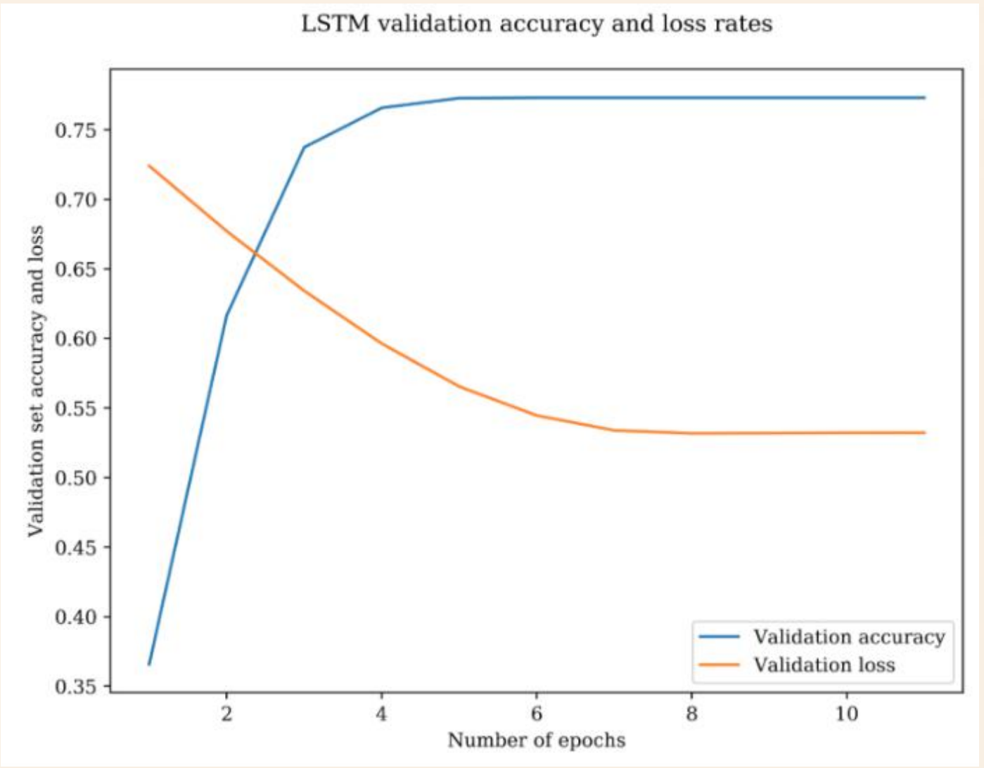


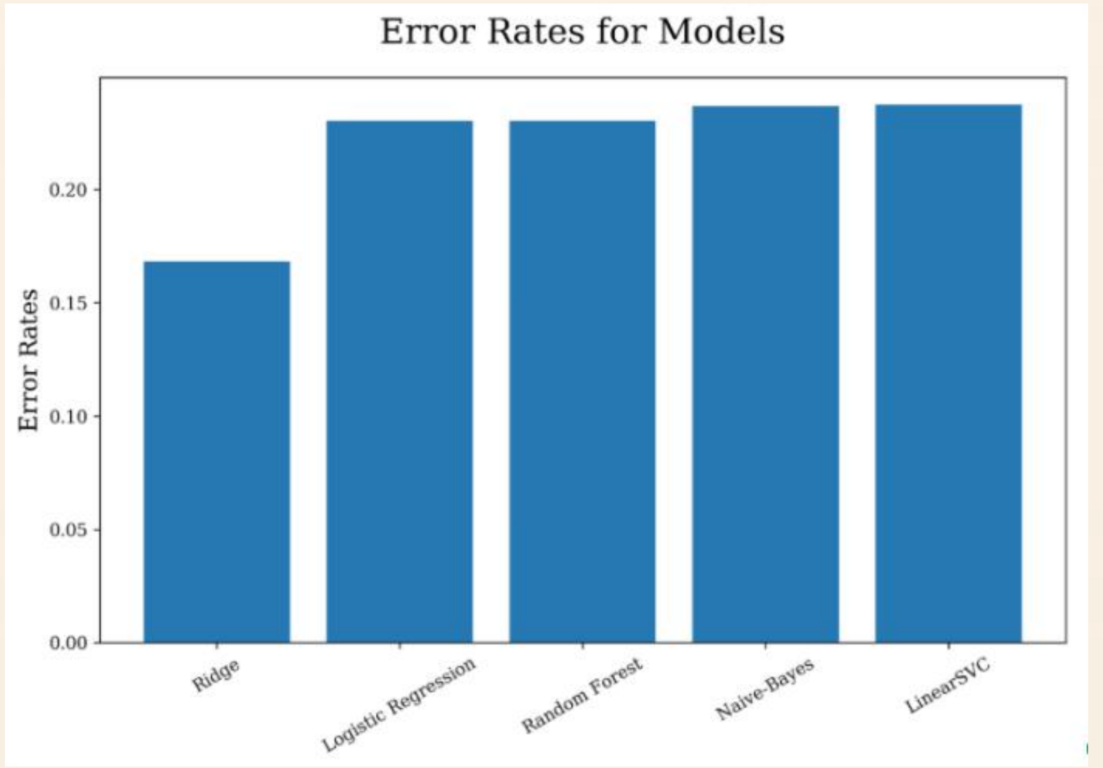
Discussion:

- ◆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 (bigrams & trigrams)

- ◆Model selection: Ridge
- ◆Feature Importance: Bigrams with a pronoun paired with a noun (pro, n), determiner (det, n) is most relevant to Age.
  - ➢ *Specific types of clauses are learned before others*
  - ➢ *More detailed linguistic analysis would be needed*
  - ➢ *Children in this data already passed acquisition stage.*

References

Gillam, R. B. & Pearson, N. (2004). Test of Narrative Language. Austin, TX: Pro-Ed Inc.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. NIPS.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2018, April 23). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System. Retrieved from psyarxiv.com/93mwx