# Pedantic Speech in Children with Autism Spectrum Disorder

Grace Lawley, CSE PhD Student

CSEE Seminar

May 14th, 2019

OHSU

# Outline

1. About me
2. Current project
3. Pedantic speech in Autism Spectrum Disorder
    a. History of “pedantic” in clinical setting
    b. Past CSLU work on pedantic speech
4. My work on pedantry
    a. Automated measurement of pedantic speech
    b. Corpus linguistics
    c. Revisiting previous approach

# About Me

BA in Mathematics

- Lewis & Clark College, 2017

CSE PhD Student

- September 2017 - Now
- Advisor: Steven Bedrick (previously Jan van Santen)

Academic Interests

- Computational Linguistics
- Natural Language Processing
- Discrete Math, Statistics
- Speech & Language Disorders

Plants

Dido

# Current Project

*Automated Measurement of Language Outcomes for Neurodevelopmental Disorders (R01DC012033)*

- Eric Fombonne, Steven Bedrick, Jill Dolata
- Joel Adams, Alexandra Salem, Heather MacFarlane

Focuses on the language of children with one of three neurodevelopmental disorders

- Autism Spectrum Disorder (ASD)
- Fragile X Syndrome (FXS)
- Down Syndrome (DS)

# Primary Goals of the Grant

1. Optimize the parameters for our *Automated Discourse Measures* (ADMs)
    - NLP algorithms developed to measure language difficulties in ASD
2. Evaluate test-retest reliability of each ADM on data collected by the UC Davis MIND Institute
    - Longitudinal data
    - ASD, FXS, and DS

# Research

# Autism Spectrum Disorder (ASD)

Core characteristics

- Restricted, repetitive interests
- Difficulties with social communication

These language difficulties can appear in many different forms…

## Idiosyncratic Language

- Using a standard word in an unexpected way
- “**schedule**” vs. “**sequence**” (Volden & Lord, 1991)

## Neologisms

- Invention of a non-word
- “**bruises**” vs. “**bloosers**” (Volden & Lord, 1991)

## Pedantic Speech

- Inappropriately formal, adult-like, overly specific
- **A hole in a sock** vs. “**a temporary loss of knitting**” (Wing, 1981)

# Pedantic Speech

- Has yet to be firmly defined
- When it is defined, definitions are vague
- Various interpretations
- *Can be pedantic in many different ways…*

## Vocabulary choice

- “I ate crustaceans for lunch”
- “I ate shrimp for lunch”

## Level of detail

- “First you need to check the expiration date, then get the can opener…”
- “Mix in the tomato paste”

# “Pedantic” in ASD Literature

## Asperger, 1944

*“He also tortured himself with his obsessive **pedantries**. For example, he had wanted a pullover for Christmas, but because this wish could not be granted, he was given a particularly nice shirt and some toys as well. He was inconsolable over this ‘incorrectness’.”*

- Uses “pedantic” to describe behavior, not speech
- “Pedantries” = “insistence on sameness” in the DSM-5

# “Pedantic” in ASD Literature

## Rutter, 1965

*“They may have **pedantic** ways of putting things, using ‘**officialese**’ as one father put it”*

*“…there was frequently a **formality** of language, a lack of ease in the use of words. The children spoke in a way one might do when learning a foreign language.”*

## Van Krevelen, 1971

*“…his vocabulary bears the mark of parliamentary or townhall [sic] language reserved **more for written than spoken address**”*

# “Pedantic” in ASD Literature

## Wing, 1981

*“The content of speech is abnormal, tending to be **pedantic** and often consisting of lengthly disquisitions on favourite subjects”*

*“…he was speaking in long, involved, **pedantic** sentences that **sounded as if they had come from books**.”*

# “Pedantic” in ASD Literature

- No standard definition of “pedantic” speech
- Typically includes speech that is
  - Inappropriately formal
  - Adult-like, overly-sophisticated
  - Lengthly, providing too much detail/information
  - More similar to written language than spoken language

*Can we create an automated method to measure “pedantry” based on these descriptions?*

# How is this currently measured?

## Autism Diagnostic Observation Schedule (ADOS)

- Standard ASD assessment tool
- Series of semi-structured, examiner led activities
- Coding scheme for behaviors characteristic of ASD

Pedantic Speech

> *"Use of words or phrases tends to be more repetitive or formal than that of most individuals at the same level of expressive language, but not obviously odd…"*

Limitations

- Subjective
- Inconsistent across examiners

*Good candidate for automation*

# Have there been previous attempts at automated detection of pedantic speech?

- Ghaziuddin & Gerstein, 1996
  - Developed a pedantic coding scheme
  - *Hand coded transcripts*
  - Analysis based on raw counts
  - Found the ASD group to have significantly higher pedantic ratings than the NASD group

# Past CSLU Work on Pedantic Speech

Prud'hommeaux, E. T., van Santen, J., Black, L. M., & Roark, B. (2010). **Automatic Detection of Idiosyncratic* Word Use in Autism Spectrum Disorders.** Presented at the International Meeting for Autism Research.

## Main Goals

- Can neologism usage and pedantic speech be used to distinguish between TD and ASD in children?
- Create automated methods for detecting neologisms and pedantic speech in children

## Data

- Transcribed ADOS sessions of 4-8 year old kids
  - Two diagnosis groups: TD and ASD (sample size not given)

* Define neologisms and pedantic speech as a subtype of idiosyncratic language

Prud'hommeaux, E. T., van Santen, J., Black, L. M., & Roark, B. (2010). **Automatic Detection of Idiosyncratic* Word Use in Autism Spectrum Disorders.** Presented at the International Meeting for Autism Research.

## Methods

- Calculate relative frequency for every word a given kid says in two corpora
    1. **Wall Street Journal (WSJ) corpus** to represent adult-like speech
    2. **Child Language Data Exchange System (CHILDES) corpus** to represent child-like speech
- *Neologisms*
    - Measure as number of Out-of-Vocabulary (OOV) words said
    - Words with a relative frequency of 0 are likely neologisms
- *Pedantic speech*
    - Child uses a lot of low-frequency WSJ words (i.e. adult-like words)

Prud'hommeaux, E. T., van Santen, J., Black, L. M., & Roark, B. (2010). **Automatic Detection of Idiosyncratic* Word Use in Autism Spectrum Disorders.** Presented at the International Meeting for Autism Research.

## Results

- *Neologisms*
  - ASD group used significantly more OOV words than the TD group
- *Pedantic Speech*
  - ASD group used significantly more low-frequency WSJ words than the TD group
  - No significant difference for the low-frequency CHILDES words

# Extending Prud’hommeaux et al., 2010

**Disclaimer: the following work was done during an internship at CSLU and the general set up was inherited from a previous project*

- We are interested in the outlier words since pedantic words will most likely be infrequent ones (Prud'hommeaux et al., 2010)
- Still use the same reference corpora
    - WSJ for adult-like speech
    - CHILDES for child-like speech
- Go further than looking at raw counts
    - Calculate a single **pedantry score** for each kid

## Data

- Transcribed ADOS sessions of 4-8 year old kids
  - *ASD Group*
    - Autism Language Normal (ALN), n = 25
    - Autism Language Impaired (ALI), n = 21
  - NASD Group
    - Typically Developing (TD), n = 43
    - Specific Language Impairment (SLI), n = 20
- All participants
  - Full-scale IQ > 90
  - MLU > 3.0

## Method

- Each kid is represented as a set of unique words (types)

$$\{w_0, \ldots, w_n\}$$

- For each word a given kid used, calculate the frequency of the word in the WSJ and CHILDES
- Transform frequencies, motivation:
  - A majority of the words will have very small frequencies that are close to zero (Zipf's law)
  - The rare words will have frequencies that are very, very close to zero

# Method

- Use Anscombe's inverse sine transformation* to stabilize variance

The similar transformation for a binomial variable **r**, with mean $m$ and total number $n$, is

$$y = \sin^{-1}\sqrt{\left(\frac{r+c}{n+2c}\right)}. \qquad (1{\cdot}4)$$

The optimum value of $c$ is $\frac{3}{8}$ if $m$ and $n-m$ are large. The variance is approximately $\frac{1}{4}(n+\frac{1}{2})^{-1}$.

- Used in previous CSLU work to compute one of the ADMs: *overall repetition ratio (ORR)*

**Quantifying Repetitive Speech in Autism Spectrum Disorders and Language Impairment**

**Jan P. H. van Santen**, **Richard W. Sproat**, and **Alison Presmanes Hill**
From the Center for Spoken Language Understanding, Oregon Health & Science University, Beaverton, Oregon (J.P.H.v.S., A.P.H.); Google New York, New York, New York (R.W.S.)

*Anscombe, F.J. (1948) The transformation of poisson, binomial and negative binomial data. *Biometrika*, 35, 246-254.

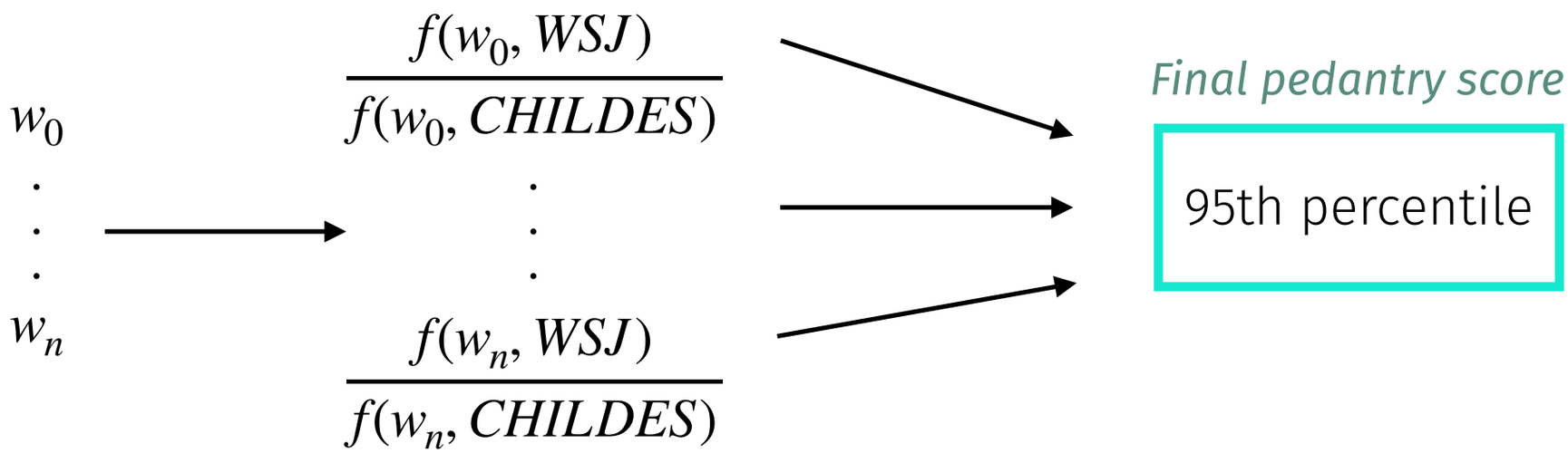
## Method

- Transformed frequencies for each word

$$f(w_i, c) = \arcsin \sqrt{\frac{count(w_i, c) + \frac{3}{8}}{total(c) + \frac{3}{4}}} \qquad \text{where} \quad w_i \in \{w_0, \ldots, w_n\}, \; c \in \{WSJ, CHILDES\}$$

- Now we have two frequency scores for each word: one for the WSJ corpus and one for the CHILDES corpus
- Combine these two values for the WSJ and CHILDES by taking the ratio
- Final overall pedantry score for each kid
  - The 95th percentile of the transformed frequencies for every unique word they said

*For each kid…*



## Analysis

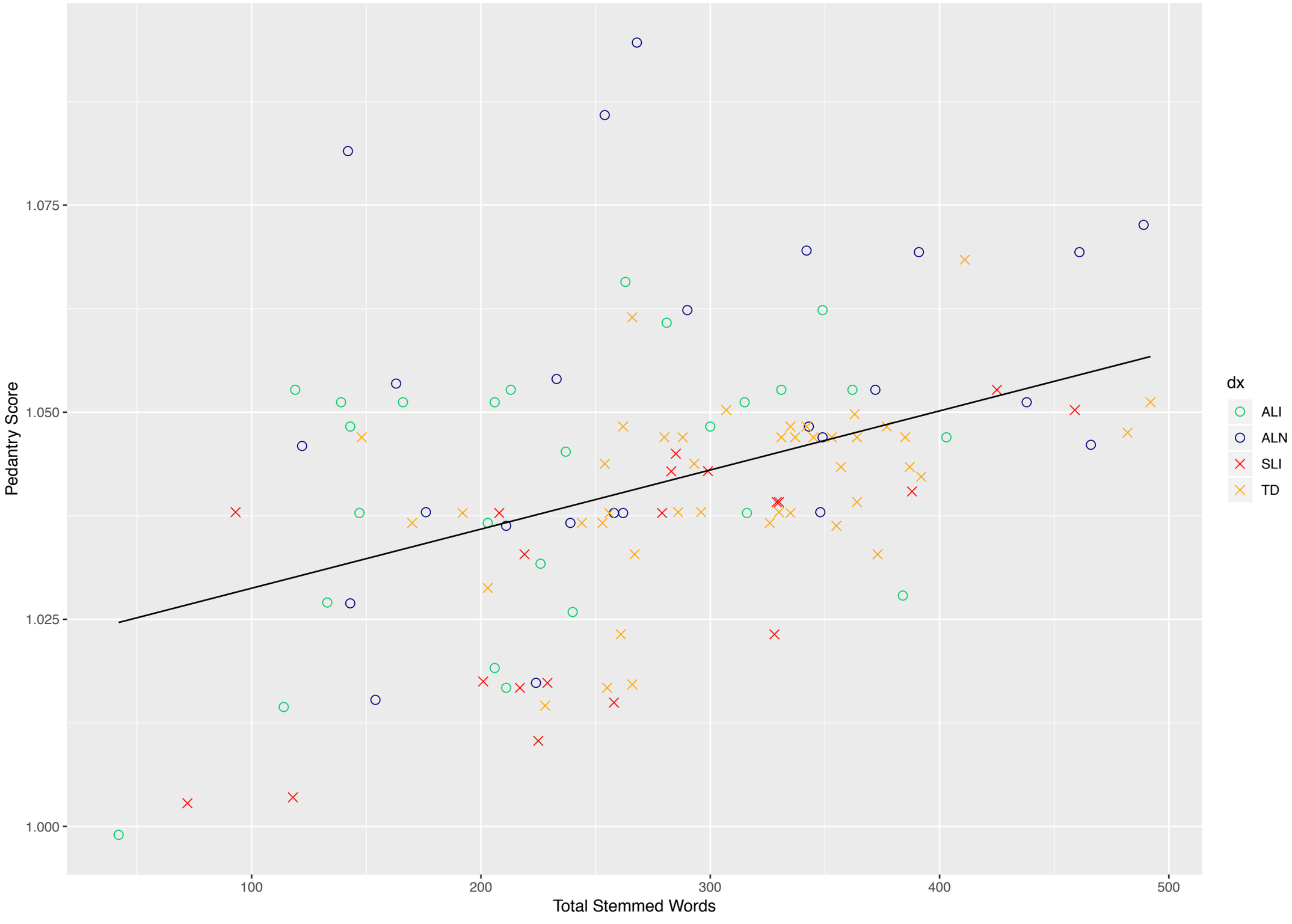
- The length of each transcript varies between kids
  - Want to compare the pedantry scores to something that is sensitive to transcript length
  - One way to capture this is by measuring lexical diversity
- For each kid, calculate the total number of unique stemmed words (using the Porter Stemmer Algorithm)

*For each kid we now have two values:*

1. A measure of overall pedantry
2. A measure of lexical diversity

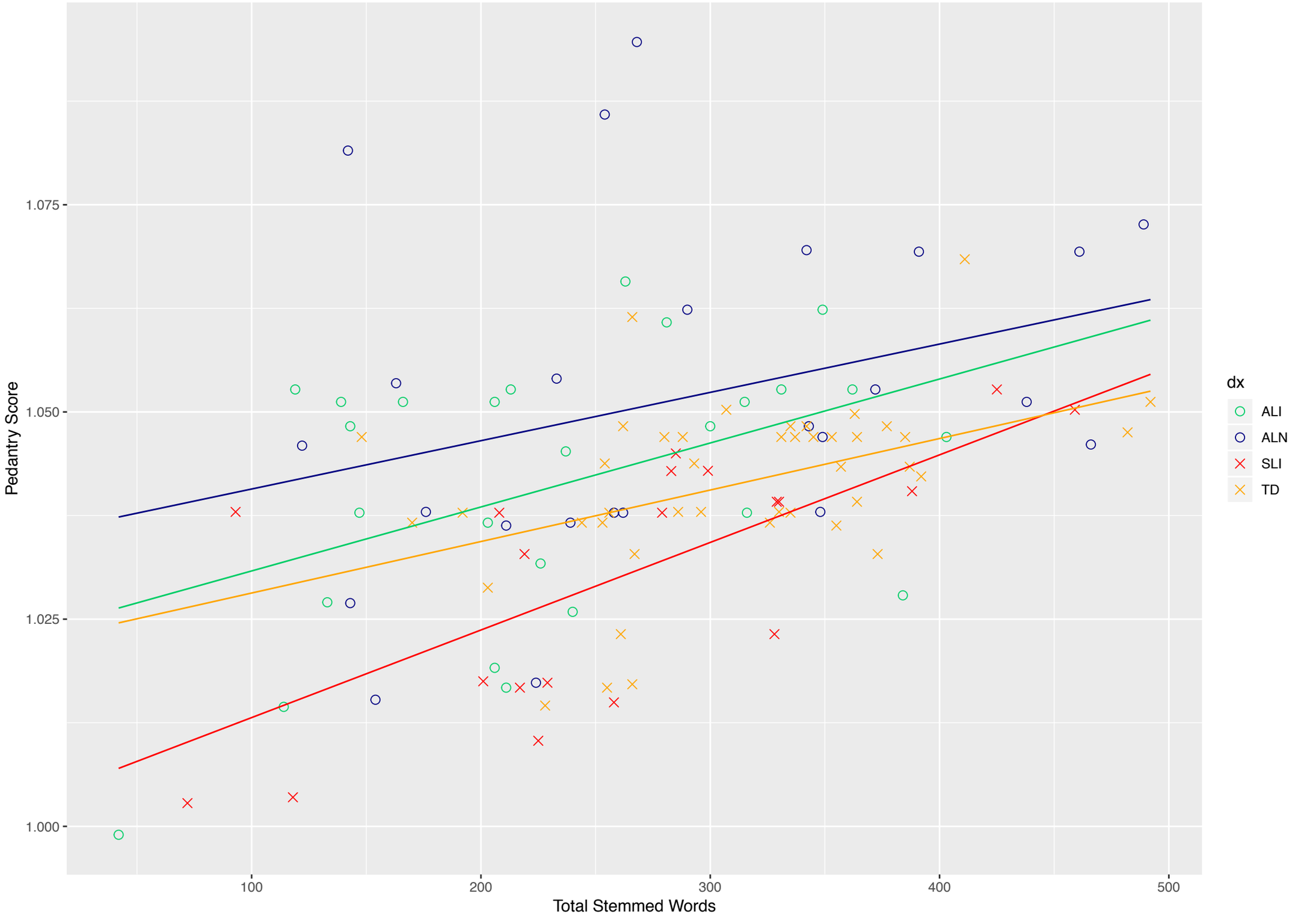
## Pedantry Scores vs. Total Stems by Child

with the line of best fit across all data points



Pedantry Scores vs. Total Stems by Child

with the line of best fit by diagnosis group



# Rethinking the Reference Corpora

## Is the WSJ corpus an appropriate proxy for adult-like language?

- *Initial thoughts*: not quite, a corpus of speech would be better
- *Thoughts now*: actually, a corpus of written language might be appropriate
  - *"…long, involved, pedantic sentences that sounded as if they had come from books" (Wing, 1981)*
- *"…his vocabulary bears the mark of parliamentary or townhall [sic] language reserved more for written than spoken address" (van Krevelen, 1971)*

Is the CHILDES corpus an appropriate proxy for child-like language?

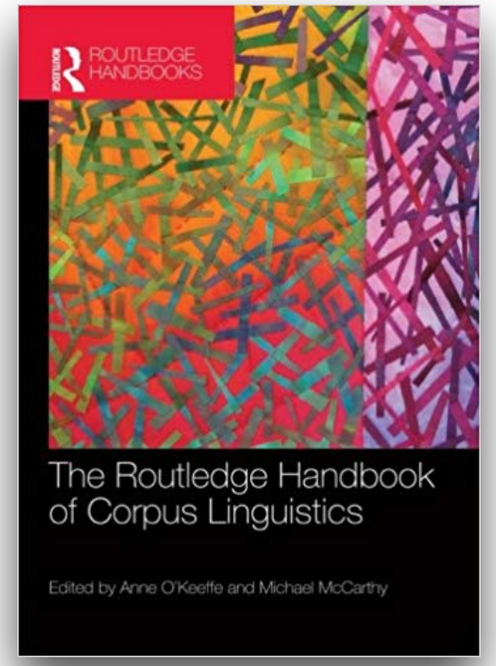
- Inherited the CHILDES corpus used from a previous CSLU project
- Consists of multiple different subcorpora from various child speech studies
- *Is it statistically sound to group these smaller corpora together into one large corpus?*

## CHILDES = Child Language Data Exchange System

- Online repository of language acquisition data
- CHILDES → Eng-NA
    - Collection of English North American Corpora
    - 54 corpora

# The Routledge Handbook of Corpus Linguistics

*Anne O'Keeffe and Michael McCarthy. Routledge, 2010*




*"…Bringing together experts in the key areas of development and change, the handbook is structured around six themes which take the reader through building and designing a corpus to using a corpus to study literature and translation."**

*https://www.routledge.com/The-Routledge-Handbook-of-Corpus-Linguistics/OKeeffe-McCarthy/p/book/9780415464895

# Inclusion/Exclusion Criteria

## *General*

- Part of the Eng-NA section of CHILDES
- Sufficient information about the details of the study and corpus is available

## *Participants*

- 3-8 years old
- English as first and primary language
- No reported gross sensory impairments (e.g. hearing impairment, congenital defects, developmental disabilities, or atypical development)
- No significant/regular exposure to another language
    - i.e. 75% or higher consistent exposure to a language other than English

# Inclusion/Exclusion Criteria

## *Language Samples*

- Naturalistic and unscripted elicitations (in either naturalistic or laboratory settings)
- Intelligible speech
- One-on-one conversations (e.g. child-examiner conversations or parent-child conversations)
- No reading from books, etc.

# Inclusion/Exclusion Criteria

## *Language Samples*

- No restricted vocabulary that is caused by the structure of the study or the experiment design
    - e.g. no samples of free play seasons for multiple participants that each involve the same set of experimenter-provided toys
- No structured speech
    - e.g. speech from an interview that has been tailored for a specific experimental interest

# 54 Corpora → 13 Corpora

1. Bloom70
2. Braunwald
3. Brown
4. Clark
5. Demetras1
6. EllisWeismer
7. Hall
8. Kuczaj
9. MacWhinney
10. Sachs
11. Suppes
12. Warren
13. Weist

# More filtering

- Filtered transcripts and utterances against the inclusion/ exclusion criteria as applicable
  - 3-8 years old
  - No reading from books, etc.
  - No restricted vocabulary

# Revisiting Pedantry Scores

# Why Revisit?

- Initial set up of groups and corpora was inherited
- Created a new version of CHILDES corpus to use to represent child-like speech
- Learned new things about text normalization along the way
- Can normalize ADOS transcripts and WSJ corpus to better align with CHILDES corpus and each other

# New Text Normalization Decisions

- Convert all letters to lowercase
- Remove all coded words - e.g. “xxx and I went to the park”
- Remove all punctuation except apostrophes
    - Keep contracts as is - e.g. “don’t” vs. “do not”
- Tokenize into unigrams

# Revisiting Measurement Approach

- Current calculation method restricts number of reference corpora to 2
  - Is the WSJ corpus enough?
- Ideas for additional reference corpora that might contain pedantic speech
  - Project Gutenberg
  - New York Times, etc.
  - Movie subtitles

# Revisiting Measurement Approach

- Currently working on a new method for measuring pedantry based on a term-document frequency
    - Gives us information about words *not* said
- What if we also calculated pedantry scores for the reference corpora?
    - Would need to define what a “participant” is
    - Easy to do for CHILDES, but what about the WSJ?
- Could then compare the pedantry scores *between* corpora

# Thank You

This work was made possible by the support and guidance of

Jan van Santen, Alison Presmanes Hill, Steven Bedrick, Jill Dolata, Kyle Gorman, Alex Salem, Rosemary Ingham, Joel Adams, & Heather MacFarlane.

*This research was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award R01DC012033*

## Hypothesis

Using infrequent/uncommon words → pedantic speech

## Approach

- Create a term-document matrix where a document corresponds to a participant
- Instead of raw counts or tf-idf values, use a transformed frequency value

$$f(w_i, c) = \log(\frac{count(w_i)}{totalwords} + 1)$$

- Explore dimensionality reduction methods to reduce dimensionality of the term-document matrix from many to 2
- Visualize results

## Limitations

- Might not capture 2+ word phrases that are pedantic
- 2 dimensions might not be enough
- Exploratory visualizations are only the first step