In case you can’t find correspondence with my mentors, we talked mostly over discord or video call

**AD lingustics**

**Logbook- Andi Liu**

2023-2024

*Recommended Schedule:*

September - Research, Send Emails and Modeling

October - Refine Project

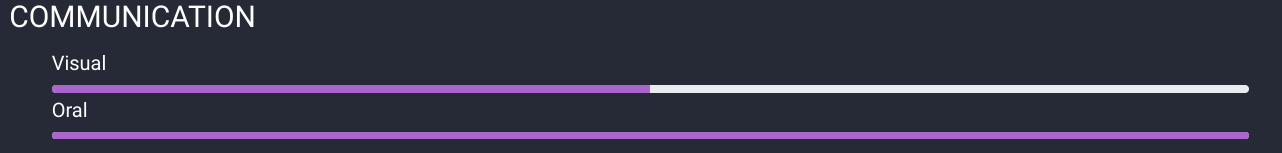
November - Start Writeup

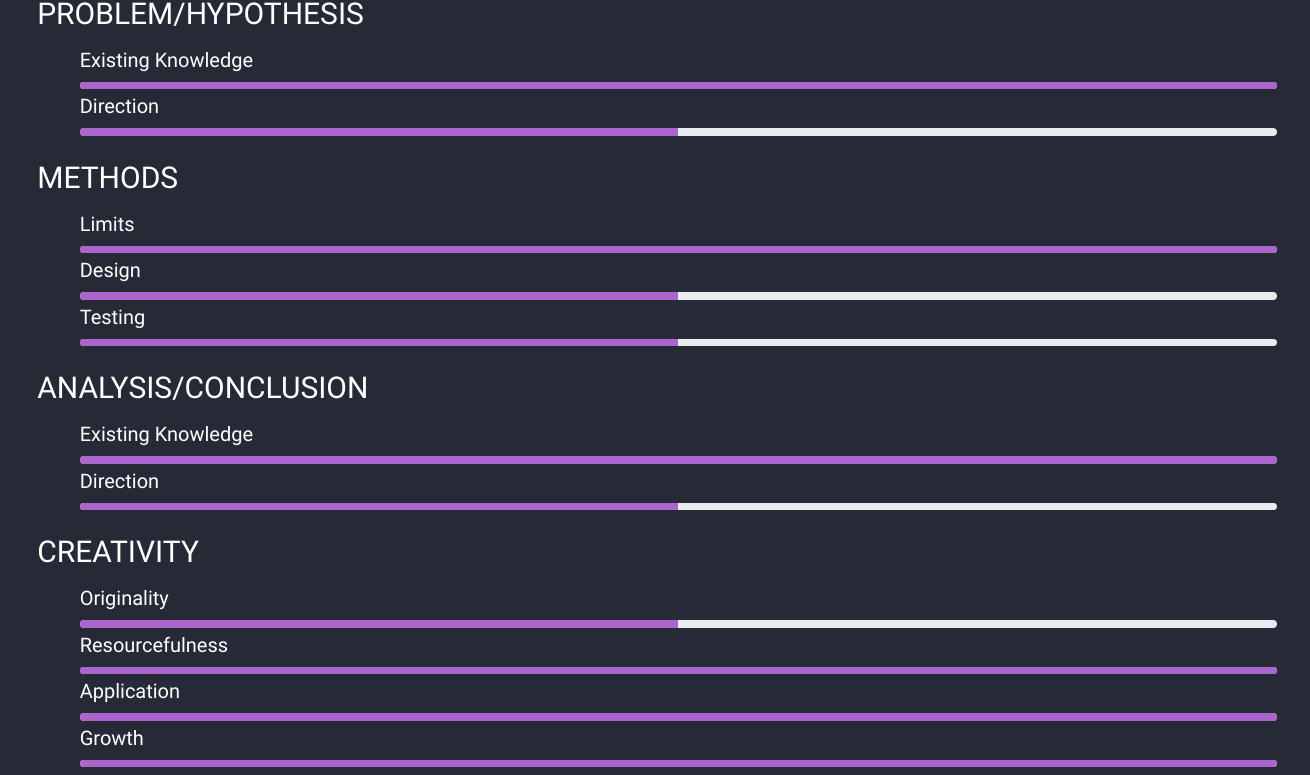
December - Finish Writeup

January - School SF

February - CYSF

-----Judging Feedback-----





**Direction (Due end of Sep)**

Problem: The field of predicting Alzheimer’s Disease from text samples is currently lacking feature sets that optimize performance and diversity from Non-European languages.

Objective(s):

1. To observe the effect of feature modification on the performance of Machine Learning models when classifying English text of varying mediums.
2. To observe the effect of feature modification on the performance of Machine Learning models when classifying Non-European (Simplified Mandarin) audio transcripts.

Hypothesis(s):

| Topic | Hypothesis |
| --- | --- |
| Stopwords |  |
| Pr |  |
| Language | Models classifying Chinese text will benefit more from |

Null hypothesis:

What is your goal?

What are you trying to achieve?

How will you achieve it?Why are you trying to achieve it?

What purpose does this project have?

Variables:

Manipulated:

* Medium - transcripts or blogs
  + Cookie, sentence, (no fluency or recall because of extreme class imbalance) - transcripts
* Adding features:
  + Addition or removal of stopwords (is Alkenani et al. right?)
  + Adding prosodic features
  + Other features:
    - Noun to verb ratio
    - Closed class to open class ratio
    - Stopword ratio
    - Idea density
    - Type-token ratio

**Resources:**

Blog Post dataset- [GitHub - vmasrani/blog\_corpus: Script to download dataset from "Detecting Dementia through Retrospective Analysis of Routine Blog Posts by Bloggers with Dementia"](https://github.com/vmasrani/blog_corpus)

Feature extraction: [Overview - CoreNLP](https://stanfordnlp.github.io/CoreNLP/)

PoS (important!!): <https://stanfordnlp.github.io/CoreNLP/pos.html>, <https://medium.com/greyatom/learning-pos-tagging-chunking-in-nlp-85f7f811a8cb>

General text classification: <https://medium.com/analytics-vidhya/nlp-tutorial-for-text-classification-in-python-8f19cd17b49e>

Using CoreNLP Step by step: <https://towardsdatascience.com/natural-language-processing-using-stanfords-corenlp-d9e64c1e1024> , <https://github.com/Ling-Ling/CoreNLP-Text-Classification/blob/master/HOWTO.md>

Stanza: <https://github.com/stanfordnlp/stanza/blob/main/stanza/models/classifiers/cnn_classifier.py>

Chinese version: <https://stanfordnlp.github.io/CoreNLP/human-languages.html>

All lang, python: <https://stanfordnlp.github.io/stanza/>, <https://medium.com/analytics-vidhya/introduction-to-stanfordnlp-an-nlp-library-for-53-languages-with-python-code-d7c3efdca118>

Step by step, en: <https://medium.com/analytics-vidhya/introduction-to-stanfordnlp-an-nlp-library-for-53-languages-with-python-code-d7c3efdca118> <https://gist.github.com/mohdsanadzakirizvi/745e150f7c15e360dc6ac2693f09f16c>

Alt english :<https://medium.com/greyatom/learning-pos-tagging-chunking-in-nlp-85f7f811a8cb>

Gen Mandarin classification: <https://www.kaggle.com/code/gpreda/chinese-text-classification>

Pos github: <https://gist.github.com/mohdsanadzakirizvi/745e150f7c15e360dc6ac2693f09f16c>

*Pitt corpus data:* [*https://dementia.talkbank.org/access/English/PItt-data.xlsx*](https://dementia.talkbank.org/access/English/PItt-data.xlsx)

## Research:

* Education and language:
* [Further education improves cognitive reserve and triggers improvement in selective cognitive functions in older adults: The Tasmanian Healthy Brain Project - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5633863/)
* [Full article: Alzheimer’s Disease Warning Signs: Gender and Education Influence Modifiable Risk Factors—A Pilot Survey Study (tandfonline.com)](https://www.tandfonline.com/doi/full/10.1080/07315724.2020.1812451)
  + Level of education influences modifiable risk factors.(ie consuming fruits and veggies, being physically active, and maintaining social relationships) The level of education has a slight correlation with a persons weekly fruit and vegetable intake (r = 0.224, p = 0.010). People with higher levels of education tend to eat more fruits and vegetables.
  + A higher level of education is associated with a healthier lifestyle and therefore less risk of developing AD

Definition Box!!

Semantic Variation - The different between the meaning or words

Open class words - Acquire new members frequently (nouns, verbs, adjectives)

Closed class - Acquire new members infrequently (pronouns, conjunctions)

* What have previous text classifiers done?
  + [Detecting Dementia from Written and Spoken Language](https://open.library.ubc.ca/media/stream/pdf/24/1.0362923/3)
    - Vaden Masrani
    - Used DementiaBack dataset
    - Used Cookie Theft Picture only for the audio transcripts
      * *Stanford Tagger* to capture various parts of speech (nouns, verbs etc) <https://nlp.stanford.edu/software/tagger.shtml>
    - Stats measured from DementiaBank:
      * Context-Free grammar rules - how often a sentence structure occurs
      * Syntactic Complexity
      * Vocabulary Richness
      * Psycholinguistics features
        + Familiarity
        + Concreteness
        + Imagability
        + Age of acquisition
        + SUTL
      * TF-IDF
      * Information Units - A list of 23 items that can be discerned from the cookie picture
      * 10-fold cross val
    - Used a blog corpus for the written portion
      * Disadvantages:
        + Text can be edited by oneself or other people
        + People with dementia may only write on good days
        + Samples may be biased to more fluent language and make it harder to tell the difference between HC and people with AD
        + They may only write topics they are familiar with
        + Topics are unconstrained
        + Observations:

Poster tended to use fewer uncommon words

Write shorter post with shorter words

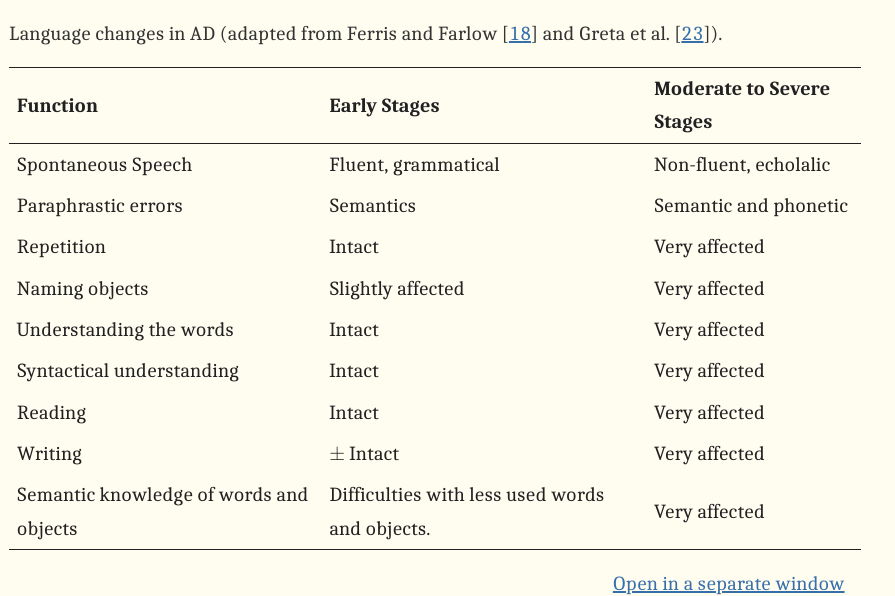
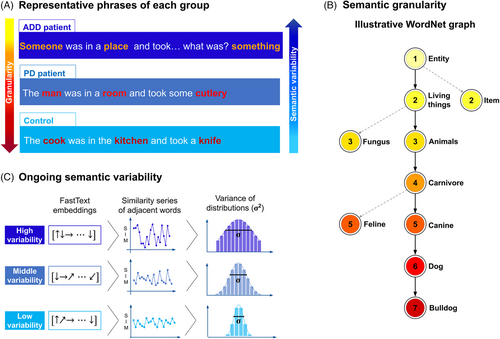
Only used 6 blogs, so finding may not be due to dementia

* + - * + Discussion:

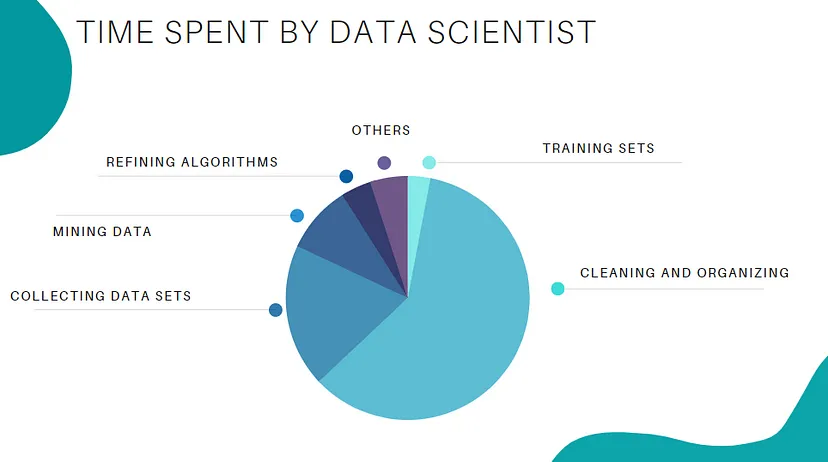
Wants to precluster the blog post dataset into certain topics more a more accurate analysis

Found that for the cookie picture, a greater decrease in performance indicates a larger feature group

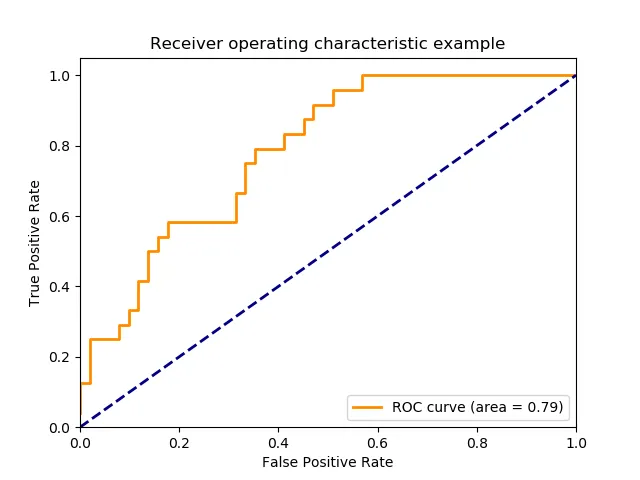
Logistic regression and random forest have better performance on average

* + <https://link.springer.com/article/10.1007/s10462-023-10419-1>
  + <https://www.sciencedirect.com/science/article/pii/S1532046421001325>
  + Measured Features:
    - Lexicosyntacitc
      * Character count
      * Word count
      * Sentence count
      * Average word length and sentence length
      * Stopwords count / functional words
      * Stopword ratio
      * Type-token count - amount of unique words
      * Type- token ratio
      * Open-class to closed class ratio
      * Idea density - total number of propositions to total word count
      * Noun to verb index
      * Verb to noun index
      * Active proposition density - ratio of verbs, adverbs and adjective to nouns
      * Passive proposition density - ratio of nouns to verbs, adverbs and adjectives
      * Word class count - absolute count of open and closed words classes
      * Word class ratio
      * Mean content density - ratio of propositions to the ration of open class to closed class
    - Method
    - Datasets preprocessing and feature extraction: In this phase, raw data from original datasets is preprocessed and transformed into a suitable format for analysis and [machine learning algorithms](https://www.sciencedirect.com/topics/computer-science/machine-learning-algorithm). This phase also involves extracting lexicosyntactic features and character-level vocabulary spaces.
    - •  
      Dimensionality reduction: This phase presents the optimization techniques of our feature spaces.
    - •  
      Classifiers training: The development of the proposed stacked fusion model involves two-level learning process; level-0 for training the single [individual classifiers](https://www.sciencedirect.com/topics/computer-science/individual-classifier) as base learners using *K*-fold cross validation approach on the training set
    - �train
    - and level-1for training the stacked fusion models on top of base learners, using the output of level-0 as the training set.
    - •  
      Alzheimer’s disease prediction: During this phase, the proposed stacked fusion models are tested using a held-out test set
    - �test
    - . This model is based on stacked generalization, which is an ensemble technique that aims to reduce the [generalization error](https://www.sciencedirect.com/topics/computer-science/generalization-error) by using the predictions of a pool of base classifiers to train another classifier based on these predictions [[31]](https://www.sciencedirect.com/science/article/pii/S1532046421001325#b0155).
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7671617/>
  + Answered 5 questions
    - Study Characteristics
      * 33 studies, 18 on AD, 9 on AD and MCI, 6 on MCI.
      * 28 focused on spontaneous speech, 7 on VF and their tasks
    - What were the characteristics of participant groups?
    - What language data were collected?
    - What features of speech and language were the most informative?
    - What methods were used to classify between groups?
    - What classification performance was achieved?
  + <https://aclanthology.org/N19-1199.pdf>
  + Only one on chinese language I can find
    - Unsupervised model
    - Transfer learning from a model training on detecting the relation between mandarin and english
  + [Speech- and Language-Based Classification of Alzheimer’s Disease: A Systematic Review - PMC](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8772820/#:~:text=Alzheimer's%20disease%20is%20characterized%20by,of%20language%20processing%20%5B18%5D).
  + Language is known to be a major problem for people with AD
    - Early stages:
      * Loss of verbal fluency,
      * Replace words with wrong and meaningless ones
      * Increased pauses
      * Breakdown in written and spoken language comprehension
    - Moderate and and Late AD
      * 
  + Biggest barriers for text and speech classification:
    - Most systems are language dependent
    - No. samples used per study are very small and the No of experiments to achieve optimal performance
    - Components may require human intervention
    - Feature sets not fully established though pitch, pauses, interruptions, speech rate can help
      * Combination of linguistic and acoustic features (duration and No. of silences,hesitation, fundamental frequency, jitters)
    - ANN should still be the base model because of their flexibility
    - Cross validation should be used to evaluate the model, and accuracy and F-score should be included in the metrics
    - Additional research needed to find optimal feature set
* What are the linguistic features of Alzheimer’s Disease?
  + <https://www.cs.toronto.edu/~kfraser/Fraser15-JAD.pdf>
  + Faber-Langendon et al found that 36% of mild AD patients and 100% of severe AD patients had aphasia
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3079403/>
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3739421/>
    - Language impairment = semantic + pragmatic impairment
      * Sematic - content of words,
  + 
  + Source ^ <https://alz-journals.onlinelibrary.wiley.com/doi/full/10.1002/dad2.12276>
  + <https://aclanthology.org/N18-2110.pdf>
    - Refound known linguistic characteristics
      * Short answers and busts of speech
      * Repeated requests for clarification
      * Interjections
  + Xinmaio Liu
    - Observed that patients used more pronouns than nouns due to noun naming difficulties
    - Referred me to paper below

Lai and Lin (2013)

* <https://hal.science/hal-02339000/document>
* One way to access language is through narrative speech
  + Can be analyzed on a number of levels
    - Phonology
    - Morphology
    - Syntax
    - Semantics
    - Pragmatics
* <https://www.frontiersin.org/articles/10.3389/fneur.2022.1025660/full>
* <https://ieeexplore.ieee.org/document/9218925#citations>
* What are the best measures for Alzheimer’s disease (MCI???)
* How do text classifiers work?
  + Two schools of thought
  + 1. Semantic-lexical disorders result from loss of information in semantic representation
    - Deterioration in content of information stored by AD individuals
  + 2. Semantic information intact but AD individuals has diffuctuly accessing the information, lexical retrieval deficits.
  + Study found that
* What is Feature Engineering and Representation Learning as a whole?
  + (<https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10> )
  + Feature Engineering creates features that weren’t in the training set data
    - Bad features will have negative impacts on the model
    - Tip! Try plotting data to looks for outliers and inconsistencies
    - **Three ways to find errors in data**
      * **Domain Knowledge - ask an expert in the field**
      * **Visualization Method - Plotting data to look for inconsistencies**
      * **Use Statistics -**
        + **Feature Creation - Create new ways/statistics to display data**
        + **Feature Extraction**
        + **EDA**
        + **Benchmark model**
    - Significance:
    - 
    - Creates optimized datasets and good models
  + Techniques:
    - Imputation:
      * Missing values are typical concerns in preparing data (Human Errors, Data flow issues, privacy concerns)
      * Fills in missing data
    - Handling Outliers
      * Remove outliers from dataset to product more accurate datasets
      * Removal - may result in a large chunk of data being removed
      * Replacing Values - the outlier can sometimes be replaced with a substitute value
      * Capping - using a value from variable distribution to replace min and max values
  + In medicine and text analysis:
  + <https://www.site.uottawa.ca/~stan/papers/1999/icml99.2_word6.pdf>
* Word embedding
  + <https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050917X00112/1-s2.0-S1877050917313480/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEGkaCXVzLWVhc3QtMSJHMEUCIQCmXW%2FbkItdN%2BkVxnBqRnDniJ4BcJSb79QJJrxT2FHq0gIgGHm7I7KbvN9lqPA3FdRQhrwPsAj4ZHIUFjyXbXrQrsoqsgUIURAFGgwwNTkwMDM1NDY4NjUiDDBpXmsK5SqwSxWZuyqPBcji2Sz6LcuAHhQfp7SAuCRHUXBE%2FDjPY99gL5TDPamaiCqZERBfmwyDe5L8YYnNoUHe0yOQRH3lA7lylloDnOlN5Qrd6ZwK9ZpmebNfwoMNiCMScw8tcCjR9%2FJoEbHUfumhuQjM05CePL6CciJJ6XiJRGBXOgXGaujaLSWah667HIqidjG%2Fxb0L0OLiKre%2FpuS0G%2F4TDkx9SZtPubrdJuu74UEWbLxwOeXUqfcVTRew99EIJi7yc947ux22QplqQ86%2F020BOmeu9SlbBbIm2sR4ljLHnnE3vKKAn8GaOkczdU2%2Fz7k%2FyFoSe7uw802Jp88yEzSyWImH%2FmLxDs%2Fhz763Ba9iajEZUD8PAP5Ori9NzqgnvFxPk3nIKM3cR0G0ceoC2fNhC0wlFcviIiugh6NRBPPB9vJAemTa0g0l0H2v13Bq99GTh65E7XNLuPF6niTrdWJTo%2BFu%2FCHNYRpdfSqqRRcg0e89f8Ud9Y%2BCEEiaz6GLBVc2SkERoCQFstpFGl1yPozo1N6%2BIJ5dT0028enquoZM5rcsBhbSMpqYsXxx9D1eV7iezW%2ByTVFDYcd9n1a5oyi6pmXixFSaSt7dCR7nhVDTQqy5v50jF1mzgguaculNqqSrpg5B2My2y9FapYP4hp0O9DECrOHlDCfCK5QU%2BRviPiU0euMlCCJfQx%2F%2Fj1bg5DmpR%2FFMMQgsWziJBBBoibhlMdqmx6yyH%2FTKF35m9Oew%2F6nVkNXbHN9uybPS1qX0zxqLfSMkxGGMlYhUVco4%2F75f8juxOYWgUE%2BhPdaZpJe9%2FCaw8hbB42CZF%2BQfxmwmKnCZkiO8pHcCmsoQ4ulQ6h0pyLNhlyJSsPEiJA5lONj6%2FyPfS%2B0eHa6Is4gwl8KOqAY6sQEWKow7543NrIXs6ik5j%2FehcgsYMAdcHUWaAyNKq5pF75RuClvSVgxGx39a7xo71hCdkYdH13igWZ3FDI8GtjNmm7Sk5CmEgW2uLKJ2xLsxzjuovdH%2FCWc5m2e%2B75DEAhnRvnUQJq1R3Wpg8WntbdOnDiLVEv5g2WfUpvtZeicOEioPBl3tPSfsSNvMpbBN1I4z2oW4BqpBjsyo9Mxe8BUsrzD1uVbkvQmdMDXZi3JVl%2FY%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20230915T001852Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYZ2SLPGGL%2F20230915%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=edec1ffcb4d07616c2d60a8cd70991a785dce6eedc0ea382bf4ab3a94df1fa99&hash=6177937b78a894f75cce54bac6684ce949e89004c6d1f6cf0571d9293ff8c11f&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050917313480&tid=spdf-a4b7cbc5-57a2-4c30-89e4-36f751bf9ea4&sid=aef254df277ca14ab20acf45f42377aa6ab5gxrqa&type=client&tsoh=d3d3LnNjaWVuY2VkaXJlY3QuY29t&ua=0007565e5707065306&rr=806cb1e72c6b6828&cc=ca>
  + <https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>
  + Most popular representation of a documents vocabulary
    - Capables of capturing a words context, semantic and syntactic similarity relation with other words
    - Word embedding are vectors of words
    - Why do we need them?
* Stats measured
* <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234#:~:text=Area%20Under%20Curve(AUC)%20is,a%20randomly%20chosen%20negative%20example>.
* Base metrics; need to know
  + IMPORTANT: “positive” refers to the patient being positive for LO-SAD. “Negative” refers to the patient being negative for LO-SAD. Also, since the classes are divided into 0 or 1, 0 can refer to a patient being 0% demented, and 1 can be interpreted as the patient being 100% demented.
  + ADDITIONAL NOTE: during training, the data is divided into X and Y. X is the data and Y are the true labels. Yhat is the predicted label
  + **TP** True Positive - Y is 1 and Yhat is 1
  + **TN** True Negative - Y is 0 and Yhat is 0
  + **FP** False Positive - Y is 0 and Yhat is 1
  + **FN** False Negative - Y is 1 and Yhat is 0

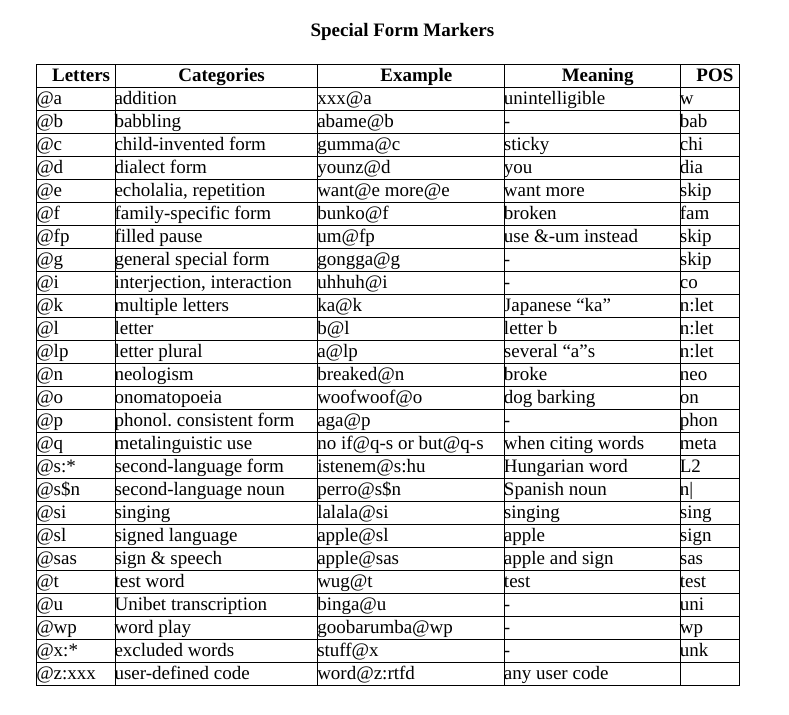
---------------------------------------------------------------------

* + Precision - Percentage of correctly classified positives over all classified positives
  + Recall - Correctly classified positives over relevant positives (all instances that *should have* been classified as positive)
* Classification accuracy
  + Only works well if both classes are balanced
    - Ex. if a dataset has 60% in class A and 30% in class B, the model can already reach 60% by classifying everything as A
  + Can inflate accuracy if the classes aren’t equal
  + Both testing and training sets should be the same proportions
* F1
  + Not all metrics are created equal
  + Measures accuracy of the model
* Area Under Curve (AUC)
  + Measure of model performance at various threshold settings
    - What number divides class 0 and 1?
  + Measures how well the model can differentiate between the two classes
  + Most widely used metrics for eval
  + **TPR** True Positive Rate - is equal to Recall, also known as Sensitivity
  + **TNR** True Negative Rate - is the proportion of correctly classified negative data points to all negative data points. It’s also know as Specificity
  + **FPR** False Positive Rate - Proportion of all negative data points classified as positive over all negative data points
  + 
  + Image: <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234#:~:text=Area%20Under%20Curve(AUC)%20is,a%20randomly%20chosen%20negative%20example>
  + TPR and FPR have values in the range [0,1]
  + AUC is the measure of the area under the ROC curve, the greater the area the high the AUC the better the model
    - When the AUC is at 0, the model is so bad that it predicts 0’s as 1’s and 1’s as 0’s, at 0.5 it is random guessing, and at 1 it is perfect

# CHAT

<https://talkbank.org/manuals/CHAT.html>

Citation: MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. 3rd edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

* Talkbank is largest open repository on spoken language data
* All data described in CHAT format
* Main line, main transcription of what the speaker said
  + Starts with an askricks, then a three letter speaker ID, a colon then a tab
  + Everything beginning on the ninth column is a series of words
  + Words - series of ASCII characters separated by spaces
    - CLAN, all character that are not punctuation markers are potentially parts of words
      * , . ; ? ! [ ] < >
      * ^ none of those can be used in words
      * Non-letter characters that can be used within words because they have special meanings
        + + \_ - @ ( )
      * Special form markers can be placed at ends of words
        + **\*SAR: I got a bingbing@c.**
    - 

* Xxx for unintelligible speech, and www for untranscribable material
* Fillers like *uh* can be coded as &-uh. In english, this can be used of all um, uh, er, eh
* Speakers may used words like &-like and &-youknow
* Different forms in other languages
* non completion text(text)text
* When word is incomplete but meaning is clear (partial omissions only)
* Will be treated as full words

**\*RAL: I been sit(ting) all day .**

* Omitted word
* Transcription in difficult and unreliable
* And 0 before word
* **\*EVE: I want 0to go.**

Anonymization

* Silencing identifying audio, names, cities, school, hospitals
* Replacing identifying forms in transcripts with
* "Lastname", "Addressname", "Cityname", and "Schoolname"
* or pseudonyms

Basic unit of CHAT transcription is the morpheme, word and the utterance

# Logbook

Aug 30, 2023

Started researching on AD linguistic features. How would I apply these features to blog posts?

Sept 4, 2023

I fond some papers that used chinese text classification, but its all unsupervised learning. Maybe I can directly use chinese instead of english. How would I find features though?

Sept 10, 2023

* Ask blake about text classification

Sept 19 2023

* Find out how to extract data from DementiaBank
* Read on text classification in mandarin

Sept 20, 2023:

* Give to Josh to review

Sept 20 and 21, 2023

* Finish all research and send emails

Oct 16

* Finished background research writeup

Oct 20

* Might modify project to determining the most significant features of AD

Nov 5

* Slight problem, no demographic data exists for the Lu corpus! Nothing!
  + I don’t even know the class or the task the participants were doing!
  + I will email the researcher of the dataset and talk to Vaden about next steps
  + [luchingching@gapp.nthu.edu.tw](mailto:luchingching@gapp.nthu.edu.tw)

Nov 11

* Starting to code
* Still can’t find Lu corpus info, might drop Chinese (consulted Vaden)

Nov 15

* I’m going to try and extract the blog post dataset today, taking a break from DementiaBank
* Some of the blogs aren’t being extracted aaa

Nov 27

* Got some advice from Vaden:
  + When testing new features for correlation, also analysis completely correlated and uncorrelated features for reference
* I found some of the blogs no longer exist. I will need to find my own

Dec 1

* Regex not working, asked Leo for help

Dec 9

* Vaden suggested I change my heatmap to non-linear, because even a little correlation is helpful.

Dec 10

* I dropped all my hypothesis expect for the TTR hypothesis

Dec 15

* Extracted the histograms today
* Added demographic information to the dataframe

Dec 16

* Ran t-test, the shapiro wilk test, and some scatter plots
  + I noticed that TTR decreases as TTC increases, in both classes and mediums
  + I ran another scatter plot to see if TTR and TTC also decreases with total word count

Jan 10

* Got both mentors to review work
* Vaden suggested restating my hypothesis to be more clear
  + Also to add less statistics to my speech

Jan 13

* Finalised my results
* The t value is lower in spontaneous speech for some reason

Jan 20

* Dr Dash suggested that a lower t-value in ss could be due to participants not doing the cookies task from memory
* Worked on tri fold printout and final writeup
* Script is a bit too long

Jan 21-25

* Finished trifold and class presentation
* Finished writeup
* The script is now 10 minutes 🙂

Jan 30

* Presented to class to get class feedback
  + Suggested I could speak louder

Jan 31

* Last minute prep
* Printed out examples of blog posts to show the judges

Feb 7

* I qualified for CYSF yay
  + Next Steps
  + Try to find a written dataset that has demographic data. If can’t find, might reconsider my project
  + Collect demographic data for spoken speech
* Going to Italy for spring break so will need to schedule a time to meet with Joshua

Feb 13

* Found demo data

The .cha id is first three numbers is the id of the participant and the number after the hyphen is the interview number

* The id numbers are not in chronological order for some reason. Its skips numbers occasionally (really annoying)
* The MMSE score in the chat file is not consistent with the one written in the demographic data

Feb 26

* Met with Dr Dash to discuss analysis

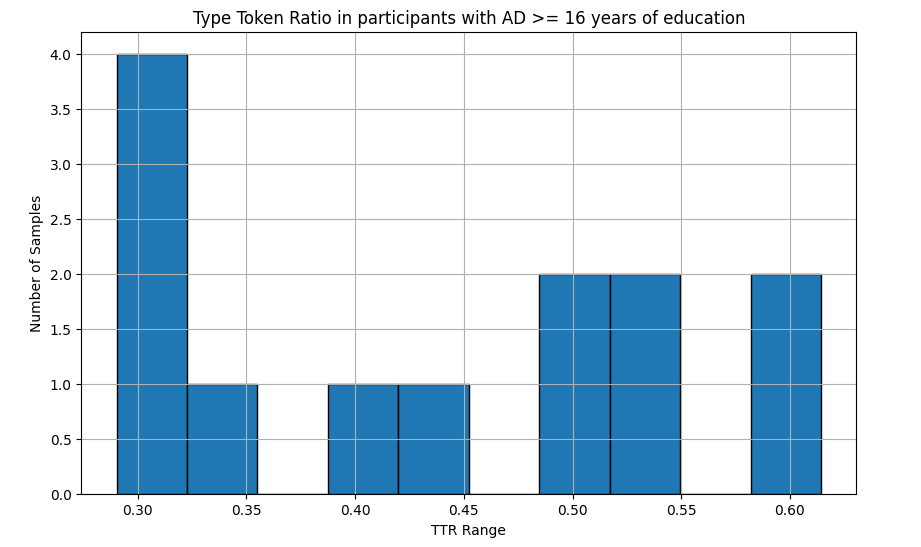
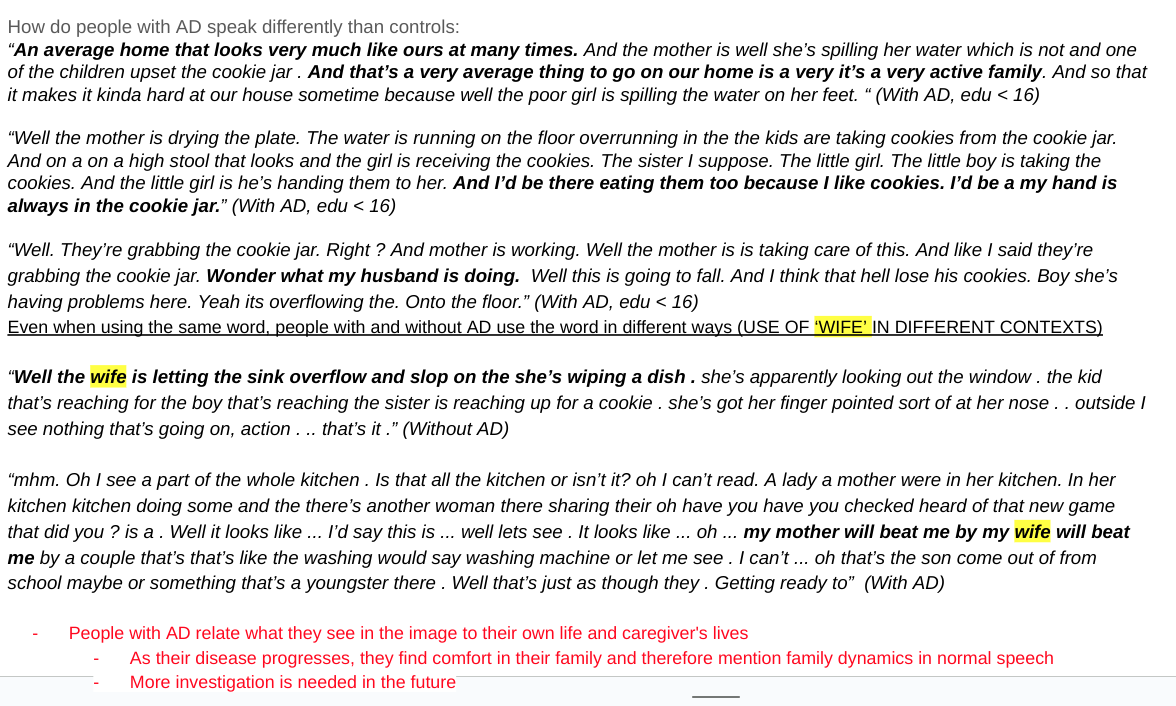
Loose plan:

* Feb - try to find dataset. If can’t find by the end of the month, might reconsider project
* March - start writeup
* April 1 - last minute plans and practice presentation
* April 12 - CYSF

Feb 27:

* I found that dividing the two groups by at the 12 year education mark only give 6 samples in the control group with less than 16 years of education
  + I will divide the groups by 16 yrs to get a more even population

March 1:

* I’m getting weird results because of splitting by education level
* 
* I don’t have many samples per group and the distribution is looking sus…
* I also notices that people with AD, regardless of education level, tend to relate what they see in the image to their own lives 

March 3:

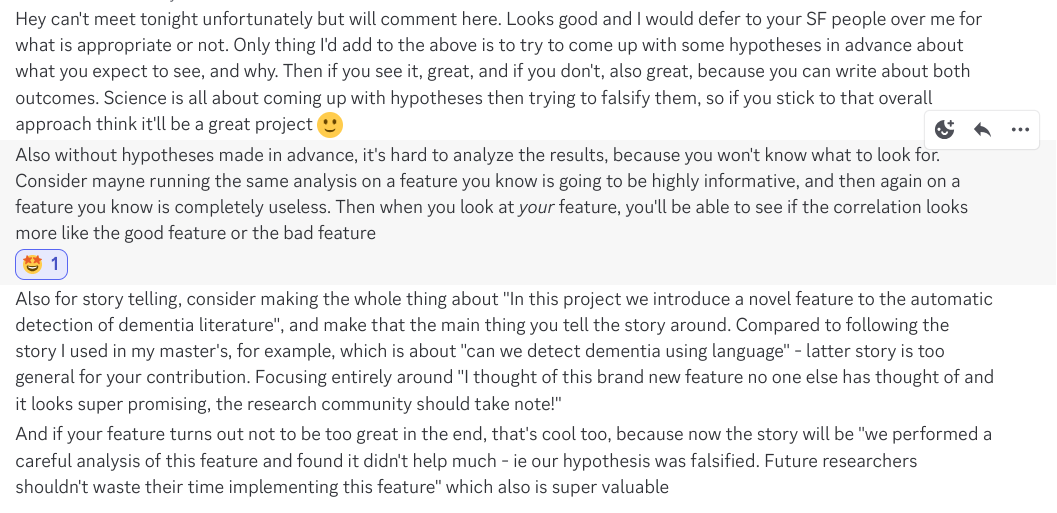
* Tanya suggested I could try something called age matching if I want to control the effects of age. It involved removing outliers, like 30 year old and 90 year olds so the mean age between the two groups is not statistically significant. This also applies to years of education
* I found this is more effective than my previous solution
* She also told me to check out verbal perseverations, which is phrase and word repetition

March 10:

* Deadline coming up
* Verbal preservations are pretty similar to echolalia, I wonder what the difference is
* I ran Cohen’s d, and the effect size is pretty small
  + Tanya says I should still include it even if its not favorable, because there are some interesting analysis I could do
* One of the papers I read said that female with AD ask more repeated questions (ie “Is the door locked?”) than men. Interesting!

March 13

* Finished Recording yey
* Posting everything on the CYSF platform

**Reviewers, thank you!**

* **~~Joshua~~**
* **Tesko?**
* **Ahat?**
* **Yohan**
* **~~Balkwill~~**
* **~~Rowley~~**
* **Fourie**
* **~~Olsen~~**
* **Jessica**

**Oct 22**

**One of my experts offered to chat with me!**

# Hypothesis

Drafts of CYSF writeup:

|  | Background research | Methodology | Data | Analysis | Conclusion and Misc |
| --- | --- | --- | --- | --- | --- |
| Draft 1 | ✔ | ✔ |  |  |  |
| Draft 2 | ✔ | ✔ | ✔ |  |  |
| Draft 2 | ✔ | ✔ | ✔ | ✔ |  |
| Draft 4 | ✔ | ✔ | ✔ | ✔ | ✔ |

**Expert questions**

* Do characteristics of language impairment (ex. More nouns than verbs, increased pauses, semantic variability) apply to Non-English language speakers? For example, would a patient speaking Chinese use more nouns like an English speaker?
  + What would be a best educated guess?
* For authors of systematic reviews: you identified ANNs as having the highest mean accuracy out of all 8 models analyzed. Do you think that the high accuracy is due to the structure of the model itself (ANNs are just better at classifying) or the high quality of other factors? (ex. Superior data)
* Simon Baker ([sb895@cam.ac.uk](mailto:sb895@cam.ac.uk))
* Professor Anna Korhonen alk23@cam.ac.uk
* Liu XinMiao E-mail: (XML); whyginger99@163.com (HYW)
* Yuefeng Li [y2.li@qut.edu.au](mailto:y2.li@qut.edu.au)
* Ines Vigo [**1170969@isep.ipp.pt**](mailto:1170969@isep.ipp.pt)

Mentor list:

* ~~Sanz~~ [~~camilasanz@gmail.com~~](mailto:camilasanz@gmail.com) ~~she/her~~
* ~~Karlekar~~ [~~sweta.karlekar@gmail.com~~](mailto:sweta.karlekar@gmail.com) ~~she/her~~
* ~~Niu~~ [~~tongn@cs.unc.edu~~](mailto:tongn@cs.unc.edu) ~~he/him~~
* ~~Bansal~~ [~~mbansal@cs.unc.edu~~](mailto:mbansal@cs.unc.edu) ~~he/him~~
* ~~Masrani~~ [~~vadmas@cs.ubc.ca~~](mailto:vadmas@cs.ubc.ca) ~~he/him~~
* ~~Haulcy~~ [~~rhaulcy@mit.edu~~](mailto:rhaulcy@mit.edu) ~~she/her~~
* ~~Li~~ [~~y2.li@qut.edu.au~~](mailto:y2.li@qut.edu.au) ~~he/him~~
* ~~Yue~~ **~~yue.xu@qut.edu.au~~** ~~she/her~~
* ~~Li~~ [~~bai@cs.toronto.edu~~](mailto:bai@cs.toronto.edu)~~. he/him~~
* ~~Hsu~~ [~~eederhsu@gmail.com~~](mailto:eederhsu@gmail.com) ~~he/him~~
* ~~Illes~~ [~~jilles@mail.ubc.ca~~](mailto:jilles@mail.ubc.ca) ~~she/her~~
* ~~Kim (DR)~~[~~mikyong@uri.edu~~](mailto:mikyong@uri.edu) ~~she/her~~
* ~~Thompson (DR)~~[**~~ckthom@northwestern.edu~~**](mailto:ckthom@northwestern.edu) **~~she/her~~**
* ~~Ribu~~ [~~ingeborgsophie.ribu@oslomet.no~~](mailto:ingeborgsophie.ribu@oslomet.no) ~~she/her~~

**Linguistics experts**

* **~~Qihui Xu (~~**[**~~q.xu@bcbl.eu~~**](mailto:q.xu@bcbl.eu)**~~) she/her~~**
* **~~Yingying Peng (~~**[**~~yingying1.peng@connect.polyu.hk~~**](mailto:yingying1.peng@connect.polyu.hk)**~~) she/her~~**
* **~~Lai~~** [**~~yhlai@nuk.edu.tw~~**](mailto:yhlai@nuk.edu.tw) **~~she/her~~**
* **~~Lin~~** [**~~ytlin@vghks.gov.tw~~**](mailto:ytlin@vghks.gov.tw) **~~Dr~~**
* **~~Castano~~** [**~~e.castano@ub.edu~~**](mailto:e.castano@ub.edu) **~~dr~~**
* **~~Dr Jeff Small~~** [**~~jeffs@audiospeech.ubc.ca~~**](mailto:jeffs@audiospeech.ubc.ca)
* **~~Dr Maxime Montembeault~~** [**~~maxime.montembeault@mcgill.ca~~**](mailto:maxime.montembeault@mcgill.ca)
* **~~Saturnino Luz~~** [**~~s.luz@ed.ac.uk~~**](mailto:s.luz@ed.ac.uk)
* **~~Corbett Aphasia Rehabilitation + Education (CARE)~~** [**~~csdcc@ualberta.ca~~**](mailto:csdcc@ualberta.ca)
* **~~Dr Elizabeth Rochon~~** ~~elizabeth.rochon@utoronto.ca~~

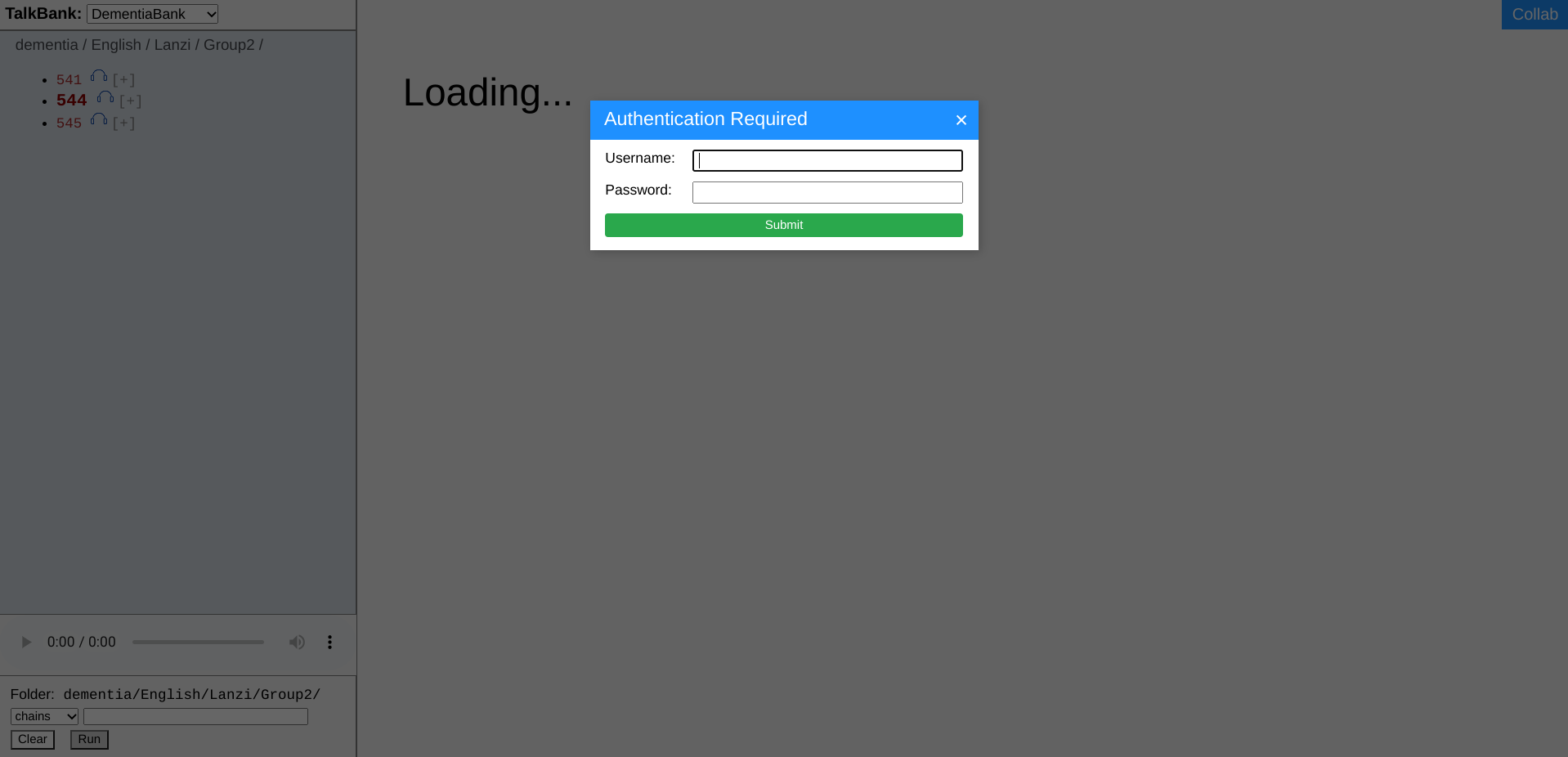
# **~~J.B. Orange jborange@uwo.ca~~**

## **~~Dr Jade CartwrightJ.Cartwright@curtin.edu.au~~**

* Tanya Dash [tdash@ualberta.ca](mailto:tdash@ualberta.ca)

*Hello,*

*My name is Andi, a grade 9 student from Canada. I am designing an audio classifier that takes in data from DementiaBank. However, I get a login prompt when I attempt to interact with the recordings in the Browsable Database. Do I need an account? If so, how would I get one? Is an account only provided in an institution?*

**

***Fig 1. The login prompt***

Thanks for your time!

Andi

---------------------------------------------------------------------------------------------------------------------------------------

**TalkBanks response 1**

**Dear Andi,**

**Thank you for your interest in DementiaBank. Unfortunately, we cannot give students direct access to these materials. You will have to ask a faculty advisor to join and take responsibility for your access. Please read the 2nd and 3rd paragraphs at the top of the DementiaBank webpage --** [**https://dementia.talkbank.org/**](https://dementia.talkbank.org/) **-- for guidance on what students should do if they are interested in doing research with this database. I should also mention that these materials are primarily intended for use in university and clinical research settings. Let me know if you have any questions.**

**Davida Fromm, Ph.D.**

**TalkBank Project**

**Psychology**

**Carnegie Mellon University**

-----------------------------------------------------------------------------------------------------------------------------------------

**Email to Stran requesting access**

Mr Stranzinger,

Thank you so much for your support on my science fair project last year! It means a lot to me.

I am looking to access an audio database called DementiaBank, which contains recordings of Alzheimer’s Disease patients for my next year's science fair project. Unfortunately, while the database is open access, it only allows faculty to oversee the use of it by students:

“*Students interested in using the data should ask their faculty advisors to join as members. The advisors must submit their own request for membership to us directly.”*

I was wondering if you could try and contact them? It’s totally fine if you aren’t comfortable doing this though, since this is a bit of a stretch.

Contact information:

* <https://dementia.talkbank.org/>
* [macw@cmu.edu](mailto:macw@cmu.edu)

Thank you for your time!

Andi

-----------------------------------------------------------------------------------------------------------------------------------------

**Requesting access, Stran**

Good day,

I am requesting membership to access the DementiaBank.

I am currently supervising a student who has been developing a science fair study of dementia over the past 3 years. It has spanned from molecular basis of demenetia to AI diagnostics.

They are interested in DementiaBank for the next step in the study project's development.

Please advise,

-----------------------------------------------------------------------------------------------------------------------------------------

**Access thread 1**

Dear Colleague,

We are happy to include you as a DementiaBank member. I am copying to Dr. Davida Fromm who will add you and send you the login information.

Good luck with your student’s project.

Best regards,

-- Brian MacWhinney Teresa Heinz Professor of Cognitive Psychology, Computational Linguistics, and Modern Languages, CMU

-----------------------------------------------------------------------------------------------------------------------------------------

**Access thread 2**

**ATTENTION: This message came from outside of the CBE and could be Spam or Phishing.**

Do not click links or open attachments unless you recognize the sender and you are certain that the content is safe. [[Learn More](https://insite.cbe.ab.ca/health-security/technology-security-access/technology-safety/Pages/phishing.aspx)]

Forward suspicious messages to phishing@cbe.ab.ca.

Dear Mark,

I'm assuming your student is the 9th grader who wrote to us a few days ago? As I explained in my reply to the student, this database is intended for use in university and clinical research settings. We have made a few exceptions and are willing to do so in your case as well. But you must understand, our research is bound by strict guidelines from our funding agencies and our Internal Review Board. I am pasting the basic membership email below, but PLEASE contact me if you have any questions.

Best,

Davida

Thank you for your interest in DementiaBank. I will add you to our membership list and send you an invitation to join our DementiaBank Google Group, which is an efficient way to communicate about relevant updates and information among members.

To access the password protected links at the DementiaBank database you need the User Name **broca** and Password **wernicke** -- all lower case letters. Please DO NOT share that access information with students, unless they are working under your supervision and understand that the materials are not to be reused or recirculated.

Don't hesitate to email with any questions about the website, the database, etc.

**Be sure to read and abide by the information about data sharing at the Ground Rules link:** <http://talkbank.org/share/> . Specifically, for the Pitt corpus, that means including this citation and acknowledging this grant support in your work:

* Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology, 51*(6), 585-594.
* NIA AG03705 and AG05133

We expect users to contribute relevant data to DementiaBank as well. When you are ready, send us an email so we can facilitate the uploading process and creation of your corpus. Finally, we encourage you to send us any posters, papers, or presentations that you produce so we can add them to our website’s links and bibliography.

Dear Dr Liu,

My name is Andi Liu, and I am a grade 9 student in Canada. For the past three years, I have been participating in my regional science fair and I have been recently looking into the role of linguistics in the detection of Alzheimer’s Disease. Your paper on sentence comprehension was very interesting to me.

<https://peerj.com/articles/8181.pdf>

I have some questions about differences between English and Chinese speakers. During your study, did you examine the differences between linguistic features of both languages? For example, a known feature of Alzheimer’s Disease in English speakers is using nouns more often than verbs, and using close class words (pronouns, etc) more than open class words (verbs, nouns etc). Does the same concept apply to Chinese speakers? Or will indicators of Alzheimer’s Disease vary across languages?

I would appreciate any advice you might have for me.

Thank you for your time,

**Response**

Dear Andi,

Thank you for your interest in my research. I only examined the sentence comprehension deficits in Chinese AD patients and did not make any cross-linguistic comparison. But I think it's interesting to make such exploration. When I collected data in the hospital, I did notice that many patients seemed to replace nouns with pronouns in oral production. They used pronouns more frequently than nouns possibly due to their naming difficulties.

Regarding the use of noun vs verb in Chinese AD, some studies did not find the same pattern as English. I suggest you read the paper by Lai and Lin(2013) for more information. I've attached the paper for your convenience.

Regards,

Xinmiao Liu

--------------------------------------------------------------------

**Email to people at cambridge**

Dear Mr. Baker/Prof. Korhonen

My name is Andi Liu, and I am a grade 9 student in Canada. For the past three years, I have been participating in my regional science fair and I have been recently looking into the role of linguistics in the detection of Alzheimer’s Disease (AD). This year, my project focuses on evaluating the performance of NLP models in predicting Alzheimer's Disease from both English and Chinese text. Your systematic review on speech and text in AD has been really insightful and helpful for my research.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7671617/>

I've got some questions about how English and Chinese speakers may differ. Since you looked into many languages in the course of your research, I thought you would be the best person to ask about this. In your review, did you make any cross-linguistic comparisons between the two? For example, in English, people with Alzheimer's tend to use more nouns than verbs and rely more on words like pronouns. Does the same happen in Chinese, or do we see different language patterns indicating AD?

Additionally, I was wondering if you would be interested in mentoring and guiding me over the course of my project. The timeline is from October to January. I would appreciate any advice you might have for me.

Thank you for your time and support.

Sincerely,

Andi Liu

--------------------------------------------------------------------------

**Questions to Prof Li**

Dear Mr Li,

My name is Andi Liu, and I am a grade 9 student in Canada. For the past three years, I have been participating in my regional science fair and I have been recently looking into the role of linguistics in the detection of Alzheimer’s Disease (AD). Your paper on detecting AD from spoken and written language has been very helpful in my research.

<https://www.sciencedirect.com/science/article/pii/S1532046421001325>

I have some questions about the use of stopwords in

--------------------------------------

Dear Ms Vigo,

My name is Andi Liu, and I am a grade 9 student in Canada. For the past three years, I have been participating in my regional science fair and I have been recently looking into the role of linguistics in the detection of Alzheimer’s Disease (AD). Your systematic review on speech and language classification in AD has been very helpful in my research.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8772820/#:~:text=Alzheimer%27s%20disease%20is%20characterized%20by,of%20language%20processing%20%5B18%5D>

I have some questions about the models used in the studies you reviewed. ANN models were shown to have one of the highest mean accuracies out of all 8 analyzed. Was the high accuracy due to the structure of the model (ANNs are inherently better at classifying AD language data) or to other factors of the individual studies themselves (ex. studies who used ANN models had access to more intensive data)?

I would appreciate any advice you might have for me. Thank you for your time and support.

Sincerely,

-------------------------------------------------------------------------------------------------------------------------------

Dear (nme),

My name is Andi and I am a grade 9 student in Alberta, Canada. For the past four years I have been participating in my regional science fair. I was doing some research on the role linguistics plays in AD diagnoses, when I came across your article.

<https://browse.arxiv.org/pdf/1804.06440.pdf>

This year, I am working on a project titled “Enhancing Alzheimer's Disease Detection through Feature Engineering in Natural Language Processing.” The goal of this project is to observe the effect that linguistic features have on the detection of AD in various mediums, such as written and spoken language.

I am writing to ask whether you would be interested in mentoring and guiding me in this project. I feel that you would be able to help me greatly as you are highly experienced in this field. The timeline of this project is from November to April.

Thank you for your time.

Sincerely,

Andi Liu

-------------------------------------------------------------------------------------------------------------------------

Dear (nme),

My name is Andi and I am a grade 9 student in Alberta, Canada. For the past four years I have been participating in my regional science fair. I was doing some research on what linguistic features are present in Alzheimer’s Disease (AD), when I came across your article.

(...)

This year, I am working on a project titled “Investigating Linguistic Features of Alzheimer’s Disease.” The goal of this project is to observe how linguistic features of AD change across different languages and mediums.

I am writing to ask whether you would be interested in mentoring and guiding me in this project. I feel that you would be able to help me greatly as you are highly experienced in this field. The timeline of this project is from November to April.

Thank you for your time.

Sincerely,

Andi Liu

---------------------------------------------------------------------------------------------------------------------------

**Response from Vanden**

Hi Andi,

I would be happy to! More than happy to assist young scientists :) Would you like to have a video chat sometime next week?

Best,

Vaden

On Fri, Oct 20, 2023 at 1:01 AM Andi Liu <andil5@educbe.ca> wrote:

| [CAUTION: Non-UBC Email] |
| --- |

--

Vaden Masrani

Senior Research Scientist, Huawei Technologies Canada

Web: [vmasrani.github.io](http://vmasrani.github.io/)

Github: <https://github.com/vmasrani>

Podcast: [incrementspodcast.com](http://incrementspodcast.com/)

**---------------------------------------------------------------------------------------------------------------**

**Hey Andi!**

**Busy on the 24th unfortunately, would next Tuesday same time work? If so send a Google meet invite to this email, if not Thursday next week works as well. Looking forward to chatting, and call me Vaden! last name is too formal :p**

**Best,**

**- V**

**Email to linguistics experts**

Dear (nme),

My name is Andi and I am a grade 9 student in Alberta, Canada. For the past four years I have been participating in my regional science fair. I was doing some research on the linguistic differences between Chinese and English when I came across your article.

<https://escholarship.org/content/qt15z844t3/qt15z844t3_noSplash_ff06ea0d32eb461639e2f4cf8fc0b4e7.pdf?t=rxyb50>

This year, I am working on a project titled “Linguistic Analysis in Alzheimer’s Disease” The goal of this project is to observe how the language of Chinese and English speaking Alzheimer’s Disease patients compare, and the cause of it.

I am writing to ask whether you would be interested in mentoring and guiding me in this project. I feel that you would be able to help me greatly as your research included cross-linguistic comparisons between Chinese and English. The timeline of this project is from November to April.

Thank you for your time.

Sincerely,

Andi Liu

Email to Lai and Lin

Dear name,

My name is Andi and I am a grade 9 student in Alberta, Canada. For the past four years I have been participating in my regional science fair. I was doing some research on linguistics in Alzheimer’s Disease (AD) when I came across your article.

<https://escholarship.org/content/qt15z844t3/qt15z844t3_noSplash_ff06ea0d32eb461639e2f4cf8fc0b4e7.pdf?t=rxyb50>

This year, I am working on a project titled “Linguistic Analysis in Alzheimer’s Disease” The goal of this project is to observe how the linguistics of AD changes across language and medium.

I am writing to ask whether you would be interested in mentoring and guiding me in this project. I feel that you would be able to help me greatly as your research includes determining linguistic features in AD patients. The timeline of this project is from November to April.

Thank you for your time.

Sincerely,

Andi Liu

**Contact info request**

Dear Dr Castano,

I hope this message finds you well. My name is Andi and I am a grade 9 student in Canada. I'm writing to request contact information for one of your students, Laura Martínez Carrizosa, regarding a mentorship for my science fair project. Any assistance you could provide would be greatly appreciated.

Thank you for your time and consideration.

Sincerely,

Andi Liu

--------------------------------------------General Email used------------------------------------------------

Dear name,

My name is Andi and I am a grade 9 student in Alberta, Canada. For the past four years I have been participating in my regional science fair studying Alzheimer's Disease (AD). I found your contact information [].

This year, I am working on a project titled “Linguistic Analysis in Alzheimer’s Disease” The goal of this project is to observe how differently language impairments are expressed in individuals with AD based on the medium.

I am writing to ask whether you would be interested in mentoring and guiding me in this project. I feel that you would be able to help me greatly as your research focuses on speech science and AD. The timeline of this project is from December to April. Any advice you have for me would be greatly appreciated.

Thank you for your time.

Sincerely,

Andi Liu

**Response from Dr Orange**

Good Morning, Andi Liu,

How wonderful that you are taking on an important topic. I applaud you for pursuing the topic.

I am retired from my academic position but remain engaged partially in research. I could provide some guidance but suggest you also seek help from other research faculty at Western University and faculty at other Canadian universities.

So, what next steps do you want me to take?

JB

**Reply**

Dear Dr Orange,

Thanks for your quick response!

My project looks at language samples from AD patients in two different formats, spontaneous speech and blog posts. The goal of my project is to determine if language impairments (ie. number of unique words used) are different between the two formats and healthy controls. I also want to see if any language impairment is highly correlated with the presence of AD. Your expertise would be helpful as I’m looking to explain the neuroscience behind my results.

Could you please review my final writeup and give feedback on it? My experiment and data collection should be completed by the end of the year. I am mostly looking for feedback on the structure and if I missed mentioning certain pieces of information. I also have several questions about language impairment in AD I was hoping you could answer.

Regards,

Andi

Response from CARE clinic

Hello Andi,

Thank you for reaching out to us.

I had forwarded your request to Communication Sciences and Disorders Department and one of our team members\* responded showing interest in your project. Her name is Tanya Dash and you are welcome to contact her at tdash@ualberta.ca.

Best wishes,

Saira

Corbett Clinic

Email to Tanya Dash

Dear Dr Dash,

My name is Andi, and I got your contact information from Saira at the Corbett clinic. Thank you for agreeing to help me!

My project looks at language samples from Alzheimer’s Disease (AD) patients in two different formats, spontaneous speech and blog posts. The goal of my project is to determine if language impairments (ie. number of unique words used) are different between the two formats and healthy controls. I also want to see if any language impairment is highly correlated with the presence of AD. Your expertise would be helpful as I’m looking to explain the neuroscience behind my findings.

Could you please review my final writeup and give feedback on it? My experiment and data collection should be completed by the end of the year. I am mostly looking for feedback on the structure and if I missed certain pieces of information. I also have several questions about language impairment in AD I was hoping you could answer.

Regards,

Andi

EMAIL FROM TANYA DASH



| | **Tanya Dash** | | --- | | 9:01 AM (3 hours ago) |  |  |
| --- | --- | --- | --- | --- |
| | to me | | --- | | | |

Hello Andi,

Thanks for reaching out.

Here are my comments and some clarification that I need:

1. I will be happy to interact with you and clear your doubts: let me know of your preference: virtual or in-person meeting? Can we schedule something together in January? Will that work? That will allow you to air your doubts!

2. do you already have participants for your project (i.e. AD and Healthy control group)? If not, do you have a plan to recruit? How do you plan to do it? Is it where you need my help? How many participants are we talking about? when you say "My experiment and data collection should be completed by the end of the year." you mean you are in the process of data collection? Is that right?

3. I will be happy to review your write up once finalized.

4. Please provide a clearer timeline of this project and where would you like my input : only theoretical (symptomatology and neurobiology of AD) or in data collection, data analysis etc also.

I think many of these questions can be easily clarified by having a quick zoom call...say 15 minutes. I am available on Thursday and Friday. After that I am on holiday until next year!

let me know if you have any other questions

Tanya Dash, PhD  
Assistant Professor, Department of Communication Sciences and Disorders  
Faculty of Rehabilitation Medicine

UNIVERSITY OF ALBERTA  
8205 114 St NW, Edmonton, AB T6G 2G4

Room number: 2-16

University of Alberta Logo

The University of Alberta respectfully acknowledges  
that we are situated on Treaty 6 territory, traditional  
lands of First Nations and Métis people.

REPLY

Dear Dr Dash

Thank you for your quick response! I am available on Friday after 1pm. How does 2:30pm sound?

Regards

Andi

-----------------------------------------------------------------------------------------------------------------------------

Dear Dr Dash,

I hope you had an enjoyable winter break! I can’t believe it's 2024 already. I know we discussed my project in December, and you suggested that I send my rough draft of my analysis. I was wondering if we could meet up later this week and discuss it (Friday?). I’m free all evenings except for Monday. The timeline for the rest of the month is as follows:

Jan 20: Analysis and Conclusion Due

Jan 29: Writeup due

Feb 2: School Science Fair

Also, while the analysis is 40 slides long, I am only using the first 35ish for my school science fair. Link: <https://docs.google.com/presentation/d/1rzNbJX33_EQLU6y7bvgtCUfpA5Hdiju9nDnKecMXusw/edit#slide=id.g2b07aa320b3_0_9>

Happy New Year!

Andi

---------------------------------------------------------------------------------------------------------------------------

Thanks Andi for reaching out.

I have a very tight schedule, I can make it at 4pm on Friday and discuss your results and interpretations. If you have any time preference (on Friday) please let me know, I will try to accommodate the request.

I acknowledge that you are also on a very tight schedule, good luck with your timeline.

best

T

Tanya Dash, PhD  
Assistant Professor, Department of Communication Sciences and Disorders  
Faculty of Rehabilitation Medicine

UNIVERSITY OF ALBERTA  
8205 114 St NW, Edmonton, AB T6G 2G4

Room number: 2-16

University of Alberta Logo

The University of Alberta respectfully acknowledges  
that we are situated on Treaty 6 territory, traditional  
lands of First Nations and Métis people.

-------------------------------------------------------------------------------------------------------------------------------

**After CYSF qualifications:**

Dear Dr Dash,

I’m pleased to inform you that I qualified for CYSF! The fair is on April 12, so I want to spend February and March applying the following modifications:

* Trying to find a written language dataset with demographic data. If I can’t find any, I may reconsider my project
* Analyze demographic data from the Pitt corpus
* I’ll try to do logistic regression if I have time, but I don’t think this is as important

I was wondering if you had any suggestions or resources regarding my next steps. Thank you so much for your assistance with my project these last few months!

Best

Andi

-------------------------------------------------------------------------------------------------------------------------------

Hi Dr Dash,

I hope you're doing well ! I have attached my analysis below. It turns out I did end up reconsidering my project a bit, and I'm analysing the level of education as a variable.

I do have some concerns about my project:

- The Pitt corpus has the race of the participant listed, but doesn't say what the numbers mean. I see that they recorded race '1' and race '2' Is there an industry standard for this? Could race 1 mean Caucasian and race 2 mean Non-Caucasian? It's one of my control variables anyway.

- As part of my analysis, I found that participants with < 16 years of education have a higher TTR and larger vocabulary than participants with >=16 years of education (this is statistically significant). This is contrary to all my research. Is this human error or is there evidence for this?

Thank you!

-Andi