“Is There Anything Else?”: Examining Administrator Influence on Linguistic Features from the Cookie Theft Picture Description Cognitive Test

Changye Li[1], Zhecheng Sheng[2], Trevor Cohen[1], and Serguei Pakhomov[2]

[1]University of Washington
[2]University of Minnesota
[1]{changyel,cohenta}@uw.edu
[2]{sheng136, pakh0002}@umn.edu

Abstract

Alzheimer’s Disease (AD) dementia is a progressive neurodegenerative disease that negatively impacts patients’ cognitive ability. Previous studies have demonstrated that changes in naturalistic language samples can be useful for early screening of AD dementia. However, the nature of language deficits often requires test administrators to use various speech elicitation techniques during spontaneous language assessments to obtain enough propositional utterances from dementia patients. This could lead to the “observer’s effect” on the downstream analysis that has not been fully investigated. Our study seeks to quantify the influence of test administrators on linguistic features in dementia assessment with two English corpora the “Cookie Theft” picture description datasets collected at different locations and test administrators show different levels of administrator involvement. Our results show that the level of test administrator involvement significantly impacts observed linguistic features in patient speech. These results suggest that many of significant linguistic features in the downstream classification task may be partially attributable to differences in the test administration practices rather than solely to participants’ cognitive status. The variations in test administrator behavior can lead to systematic biases in linguistic data, potentially confounding research outcomes and clinical assessments. Our study suggests that there is a need for a more standardized test administration protocol in the development of responsible clinical speech analytics frameworks.[1]

1 Introduction

Alzheimer’s Disease (AD) dementia is a neurodegenerative disease that causes progressive decline in cognitive function. Even though AD currently has no cure, a timely diagnosis is imperative to alleviate negative consequences of delayed or absent diagnosis including emergency events, family strife, and exposure to scam artists praying on the vulnerable (Stokes et al., 2015). Changes in naturalistic language samples collected from individuals at high-risk for dementia have been identified as one of the early signs of AD (Almor et al., 1999; Blanken et al., 1987; Bucks et al., 2000), showing its potential as an early screening tool. However, analyzing speech samples is labor-intensive and time-consuming. Contemporary studies predominately focus on automated prediction and detection of such changes with language models with considerable success in distinguishing the speech of dementia patients and healthy controls (for recent reviews, see Shi et al. (2023); Ding et al. (2024)). Despite these advances, this line of research often faces the limited data availability. As noted in Shi et al. (2023), the majority of prior work focuses on analyzing naturalistic speech samples using the transcripts of “Cookie Theft” picture description cognitive task produced by English-speaking cohorts in the Pitt corpus (Becker et al., 1994).

While several prior studies have focused on connected speech from non-English speaking participants (e.g., French (Rousseaux et al., 2010b), Spanish (Custodio et al., 2020), and German (Weiner et al., 2016)), a very limited discussion has been held in prior literature on the influence of test administrators. Similarly, methods for data collection, such as optimal sample duration, distance to the microphone, and presence of background noise, have not been standardized (Seyed Ahmad Sajjadi and Nestor, 2012). In addition, the impaired communication ability of people with dementia (Ash et al., 2006; Hier et al., 1985; Rousseaux et al., 2010a) creates additional barriers for their caregivers (Eggenberger et al., 2012; Banovic et al., 2018). This could also extend to neuropsychological assessment batteries such as picture description tasks, which are used extensively by speech-

[1] Our code is available at `https://github.com/LinguisticAnomalies/turns`

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 91–103
May 3, 2025 ©2025 Association for Computational Linguistics

language pathologists in the management of clients with language disorders, including aphasia and dementia (Cummings, 2019; Berube et al., 2019). Prior works have demonstrated that test administrators often perform a variety of speech elicitation techniques to extract additional propositions from aphasic patients (Menn and Obler, 1989; Caplan and Hanna, 1998). As a number of studies have argued in favor of a similarity of linguistic behavior in patients with dementia and aphasia (Gewirth et al., 1984; Nicholas et al., 1985; Blanken et al., 1987; Gumus et al., 2024), similar elicitation strategies may be employed when collecting speech samples from dementia patients. This could lead to the "observer effect" (Labov, 1973) in feature values as many distinct linguistic features are sensitive to the length of the text sample. A previous study (Petti et al., 2023) demonstrated that sample length is important for extracting the various language features of AD by analyzing the speech samples (e.g., public interviews, talk shows and public speeches) from cognitively healthy public figures and those diagnosed with AD dementia. However, this previous study did not address the influence of interviewers and their speech elicitation techniques on collected speech. The impact of test administrators/interviewers and the resulting reliability of linguistic features in clinical settings also remains understudied. This less-discussed gap is particularly concerning given the potential for these factors to introduce systematic biases in the assessment of cognitive decline.

To address this limitation, our study seeks to quantify the influence of test administrators on speech collected with the "Cookie Theft" picture description task. Specifically, we analyze the quantity and distribution of part-of-speech (POS) tags in task transcripts collected from participants residing in two distinct United States locations: Pennsylvania and Wisconsin. We anticipate that test administrators employ significantly more interactions to elicit speech from dementia patients compared to healthy controls, which may contribute to patients with dementia producing linguistic patterns found to be associated with dementia, such as increased use of repetitions (Hier et al., 1985), higher pronoun usage (Almor et al., 1999), and elevated lexical frequency (Bucks et al., 2000) when compared to healthy controls. We analyze the Pitt corpus and the Wisconsin Longitudinal Study (WLS) (Herd et al., 2014) datasets from the Dementia Bank. Both employ the "Cookie Theft" picture description task from the Boston Diagnositc Aphasia Examination (Goodglass and Kaplan, 1983). We aim to quantify the extent to which the linguistic features commonly attributed to dementia patients may be artifacts of the data collection and test administration process.

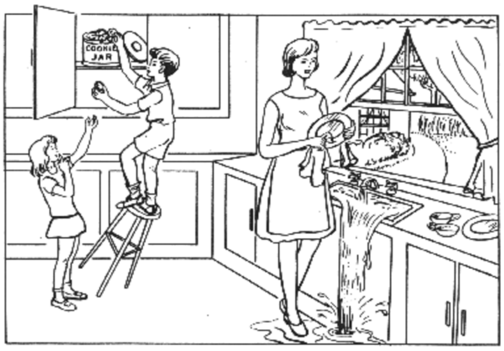


Figure 1: The "Cookie Theft" picture description stimuli.

The contributions of this work can be summarized as follows: a) we examine patterns in how test administrator involvement may relate to linguistic features observed in patient speech and their association with dementia vs. control classification; and b) our analyses raise questions about how variations in test administrator behavior might interact with linguistic patterns in clinical assessments. These observations point to opportunities for future research to investigate the role of test administration in linguistic analyses and clinical assessments.

2 Related Work

Verbal production tasks are common neuropsychological assessments for measuring language and executive retrieval functions, with the category fluency task being one of the most widely utilized in clinical settings. In this task, participants are asked to generate exemplars of specific semantic categories – such as animals or food – in a given time. While the category fluency task has demonstrated the diagnostic utility for AD screening (Monsch et al., 1992; Cerhan et al., 2002), these assessments are typically conducted in controlled clinical settings and often require longitudinal observation before a final diagnosis can be made. Such controlled testing environments can be insensitive to naturalistic language patterns (Sabat, 1994) and may miss early signs of linguistic deficits that manifest in daily communications (Crockford and Lesser, 1994). In contrast, spontaneous speech has proven to be a valuable source of information for assessing an individual's cognitive state (Bucks et al., 2000).

The “Cookie Theft” picture description task (Figure 1) is designed to elicit speech samples in pathological cohorts. Participants are asked to describe everything they observe in a picture where two children collaborate to secretly take cookies from a high cupboard shelf, while their mother is preoccupied washing dishes. Previous studies using *statistical* analyses have demonstrated many linguistic anomalies associated with AD progression, such as increased use of repetitions (Hier et al., 1985), higher pronoun usage (Almor et al., 1999), and elevated lexical frequency (Bucks et al., 2000; Cummings, 2019) compared to healthy controls. Supervised machine/deep learning methods, including transformer-based (Vaswani et al., 2017) neural language models can learn to distinguish subtle linguistic characteristics between dementia patients and healthy controls with impressive classification performance (for a review , see Ding et al. (2024)). However, such models bring an additional challenge – often the best-performing models (i.e., neural language models) are least transparent, and the less-accurate models (i.e., statistical models) are easier to explain. Limited interpretability could obscure the bias, which is particularly concerning in clinical artificial intelligence development (Reddy, 2022).

Building upon the previous findings that longer speech is important to extract distinguishable linguistic features (Petti et al., 2023) and interaction patterns between speakers are predictive of the downstream classification task (Farzana and Parde, 2022), we build statistical models to investigate the role of test administrator behavior in the manifestation of linguistic markers associated with dementia. We show that the level of test administrator’s engagement significantly impacts the linguistic features observed in the patients’ speech.

3 Method

3.1 Data

We use two publicly available datasets resulting from deploying the “Cookie Theft” picture description task during data collection: a) the Pitt corpus[2] and b) the WLS[3] corpus. The Pitt corpus includes recordings and corresponding transcripts from 319 participants. 102 out of 319 participants were classified as control subjects and 204 participants as patients categorized with any AD-related label. Specifically, we restricted the original Pitt corpus to a subset of 169 patients with an assignment of probable AD dementia and 99 healthy controls, resulting in 214 and 182 transcripts for AD patients and healthy controls, respectively.

The WLS is a large-scale, extended longitudinal study of a random sample of 10,317 men and women who graduated from Wisconsin high schools in 1957. The WLS participants were interviewed up to 6 times between 1957 and 2011. Several nueropsychological tests, including letter fluency task and category fluency task were administered in both 2004 and 2011. The “Cookie Theft” picture description task was introduced in 2011. While the WLS participants were interviewed with Telephone Interview for Cognitive Status-modified (TICS-m) for a clinical proxy diagnosis in 2020, we decide to follow a prior study (Guo et al., 2021) to build a “noisy” label with statistically determined age- and education-adjusted thresholds of 16, 14, and 12 for participants in $<$ 60, 60-79, and $>$ 79 age ranges for the category fluency score, respectively. This addresses a critical temporal aspect in AD assessment, particularly given the 9-year gap between speech data collection and clinical assessment in the WLS dataset, contrasting with the Pitt corpus where participants were diagnosed at the time of speech collection. In supporting this approach, the category fluency task, administered concurrently with the “Cookie Theft” picture description task in the WLS corpus, has demonstrated the diagnostic utility on discriminating AD patients and healthy controls, with sensitivity of 0.88 and specificity of 0.96 (Canning et al., 2004). Additionally, the number of WLS participants who completed both the cognitive tests and follow-up clinical interview remained particularly small ($<$ 35 labeled dementia patients), potentially limiting the statistical power of our study.

As a result, we restrict the original WLS dataset to a total of 1,169 participants (1,017 healthy controls and 152 dementia cased patients) who a) agreed to participant in the “Cookie Theft” picture description task and category fluency test in 2011; b) had not been diagnosed with a mental illness at the time of interview; and c) did not previously have a stroke at the time of the interview. Given the fact that the Pitt corpus contains dementia labels obtained from clinical assessments conducted concurrently with the picture description task, we

[2] https://dementia.talkbank.org/access/English/Pitt.html

[3] https://dementia.talkbank.org/access/English/WLS.html

Characteristics		Pitt		WLS	
		Control	Dementia	Control	Dementia
Gender (%)	Female	57 (59.4)	99 (68.3)	523 (51.4)	63 (41.4)
	Male	39 (40.6)	46 (31.7)	494 (48.6)	89 (58.6)
# of transcripts		182	214	1017	152
Age (mean (SD))		64.1 (7.9)	71.5 (8.63)	70.30 (4.14)	70.20 (5.75)
Education (mean (SD))		13.9 (2.4)	12.3 (2.8)	13.77 (3.01)	12.64 (2.16)

Table 1: Basic characteristics of the Pitt corpus and the WLS corpus before propensity score matching.

consider this to be an example of dementia *detection*. In contrast, the WLS dataset represents the case of dementia *prediction*. Data characteristics are provided in Table 1.

3.2 Preprocessing

We perform transcript pre-processing using TRESTLE (**T**oolkit for **R**eproducible **E**xecution of **S**peech **T**ext and **L**anguage **E**xperiments) (Li et al., 2023) for both participants and test administrators. Specifically, we remove non-ASCII characters, unintelligible words, and non-speech artifacts event descriptions or gestures. We also retain the utterances from participants in a relatively "raw" state, in which we preserve repetitions, invited interruptions, and speech repairs (self-revisions).

3.3 Topics Analysis

We segment the utterances from test administrators into individual sentences and remove the duplicates to establish a clean dataset for analysis. These utterances are then clustered based on frequency in each diagnostic group to understand the predominant conversation topics.

3.4 Linguistic feature extraction

Following the established evidence (Bucks et al., 2000; Almor et al., 1999; Hier et al., 1985; Cummings, 2019; Blanken et al., 1987), we focus our the analysis of part-of-speech (POS) tags, lexical frequency (LF), and type-to-token ratio (TTR) on utterances from participants in the Pitt and WLS corpora. We extract the counts of each POS tag for each transcript using spaCy[4] with RoBERTa (Liu et al., 2019) as the base model[5]. The log LF of each transcript is calculated using the SUBTLEX$_{\text{us}}$ corpus (Brysbaert and New, 2009). Tokens that do not appear in the SUBTLEX$_{\text{us}}$ corpus are removed as out-of-vocabulary items. TTR quantifies lexical diversity in speech samples, calculated as the proportion of unique words to total words in the transcript. We also count the number of clauses in each transcript. In this study, we define a clause as a syntactic unit centered around a verb that expresses a proposition. As a proxy of syntactic complexity (Caplan and Hanna, 1998), clause count has been shown to be a sensitive linguistic feature for detecting dementia from spoken samples (Seyed Ahmad Sajjadi and Nestor, 2012; Pakhomov et al., 2011).

Additionally, we define *turn* as the number of utterances from either participants (denoted as `par_turns`) or test administrators (denoted as `inv_turns`) in each transcript. We extract the number of turns from test administrators from transcripts for follow-up propensity score matching (PSM).

3.5 Propensity score matching

Propensity score matching (PSM) (Austin, 2011) is a statistical matching method to estimate the effect of a treatment by accounting for the covariates that predict receiving the treatment. PSM assigns a propensity score, which is the probability of treatment assignment conditional on the observed covariates. This conditional probability, serving as a balancing score, matches each individual in the treatment group to an individual in the control group in controlled experiments.

Luz et al. (2020) introduces the AD Recognition through Spontaneous Speech (ADReSS) Challenge, providing researchers with the first available benchmark that is acoustically pre-processed and balanced in terms of age and gender, both of which are risk factors for AD (Ruitenberg et al., 2001; van der Flier and Scheltens, 2005). However, it does not take into account the following possible confounding factors: a) education level, (lower education level is a risk factor of dementia later in

[4] https://spacy.io/

[5] See Table 3 in Appendix for the full list of POS tags analyzed in this study.

life and contributes to the lower linguistic ability) (Snowdon et al., 1996; Ngandu et al., 2007; Nguyen et al., 2016; Caamaño-Isorna et al., 2006); and b) the influence of test administrators, who may perform a variety of speech elicitation techniques to extract enough propositions from patients (Menn and Obler, 1989; Caplan and Hanna, 1998) in a constrained task, such as the "Cookie Theft" picture description task.

To address these concerns, we match the Pitt and the WLS corpora on: a) years of education received, and b) the number of turns from test administrators using PSM. This resulted in a balanced Pitt corpus with 167 transcripts for both dementia patients and healthy controls, and a balanced WLS corpus containing 152 transcripts for both dementia patients and healthy controls.

3.6 Statistical models

We apply z-score normalization on the POS tags, lexical frequency and TTR extracted from each transcript and treat the number of turns from test administrators as the random effects. We split the original and the matched Pitt corpus into 70/30 training/test split. We fit a generalized linear mixed models on the *matched* Pitt training split where we treat the number of turns from test administrators (inv_turns) as random effects. Our preliminary results show that fitting such a model for *matched* WLS data results in singularity (i.e., the random effects of inv_turns variance-covariance matrix is of *less than full rank*). Therefore we decide to fit generalized linear model on the WLS corpus. In addition, we compare the interaction model (models with interaction terms between inv_turns and each linguistic feature) and naïve models (models without interaction terms) and apply backward selection using Akaike's Information Criteria (AIC) (Akaike, 1998). AIC is an information-theoretic approach that estimates the distance between candidate models and the true model on a log-scale, which selects a parsimonious approximating model for the observed data. Our preliminary results show that interaction models achieve better fit with lower AIC. We then continue our analysis with the resulting interaction model for Pitt corpus ($\mathcal{M}_{\text{pitt}}$) and WLS corpus ($\mathcal{M}_{\text{wls}}$).

We also perform cross validation on each dataset to test for internal validity. Specifically, we assess the classification performance of $\mathcal{M}_{\text{pitt}}$ on both the matched Pitt test split and the matched WLS corpus and $\mathcal{M}_{\text{wls}}$ on the matched Pitt test split, respectively.

4 Results

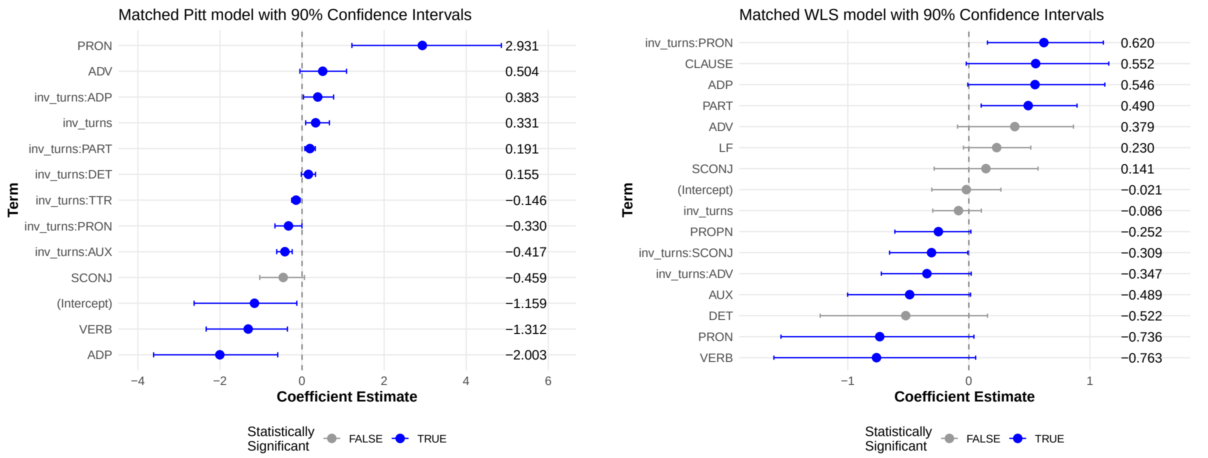
The results of PSM for the Pitt and the WLS corpus can be found in Table 4 and Table 5 in Appendix, respectively. We observed that many linguistic features preserved imbalance even after PSM, with standardized mean difference (SMD) > 0.1 (Zhang et al., 2019). It should be noted that SMD does not indicate the differences in the direction of the scale (Chandler et al., 2019) (i.e., cannot substitute the p-value from significance testing). We also observed that the WLS participants obtained a higher level of education than the Pitt participants (one-sided Wilcoxon rank sum test p-value < 0.001). These observations suggest that additional, potentially unaccounted-for variability may be influencing the results. Thus we proceeded with further quantitative and qualitative analyses.

4.1 Topics analysis

We found that test administrators' utterances usually cover the following topics: a) initiation of the task (e.g., "and there's a picture" and "what's going on in this picture"); b) acknowledgment of progress (e.g., "okay"); c) speech elicitation (e.g., "anything else", "if you see anything else" and "is there anything else"); and d) ending the task (e.g., "alright", "thank you", "that's fine" and "good"). For the Pitt corpus, test administrators said "anything else?" more frequently to dementia patients (18 times) than to healthy controls (10 times). In contrast, the WLS test administrators used the same level of speech elicitation for both groups (dementia patients: 2 times; healthy controls: 2 times).

4.2 Test administrator interaction styles

We observed a moderate negative correlation (Spearman's $\rho = -0.481$) between the number of turns used by Pitt test administrators and participants' Mini-Mental State Examination (MMSE) scores. Pitt test administrators interacted more with dementia patients who had lower MMSE scores, likely in an effort to elicit sufficient speech for analysis. As shown in Table 2, Pitt test administrators used 3 more turns on dementia patients compared to healthy controls whereas the WLS test administrators uses similar number of turns on both diagnostic groups.



(a) The coefficients of $\mathcal{M}_{\text{pitt}}$ after backward selection with AIC (b) The coefficients of $\mathcal{M}_{\text{wls}}$ after backward selection with AIC

Figure 2: The estimated coefficients and the corresponding 90% confidence intervals of $\mathcal{M}_{\text{pitt}}$ and $\mathcal{M}_{\text{wls}}$. The blue points and ranges indicate that the confidence interval does not cross zero, suggesting the estimate is statistically significant, whereas the dark gray points and ranges indicate that the confidence interval crosses zero, suggesting the estimate is not statistically significant.

Dataset/Condition			Participants' turns (mean (SD))	Test administrators' turns (mean (SD))
Pitt	Before matching	Control	13.55 (6.04)	3.16 (1.77)
		Dementia	13.54 (6.98)	6.10 (4.48)
	After matching	Control	13.44 (5.97)	3.34 (1.73)
		Dementia	12.38 (5.60)	4.38 (1.85)
WLS	Before matching	Control	14.39 (7.91)	0.75 (1.53)
		Dementia	11.97 (7.04)	0.82 (1.79)
	After match	Control	13.80 (7.76)	0.82 (1.62)
		Dementia	11.97 (7.04)	0.82 (1.79)

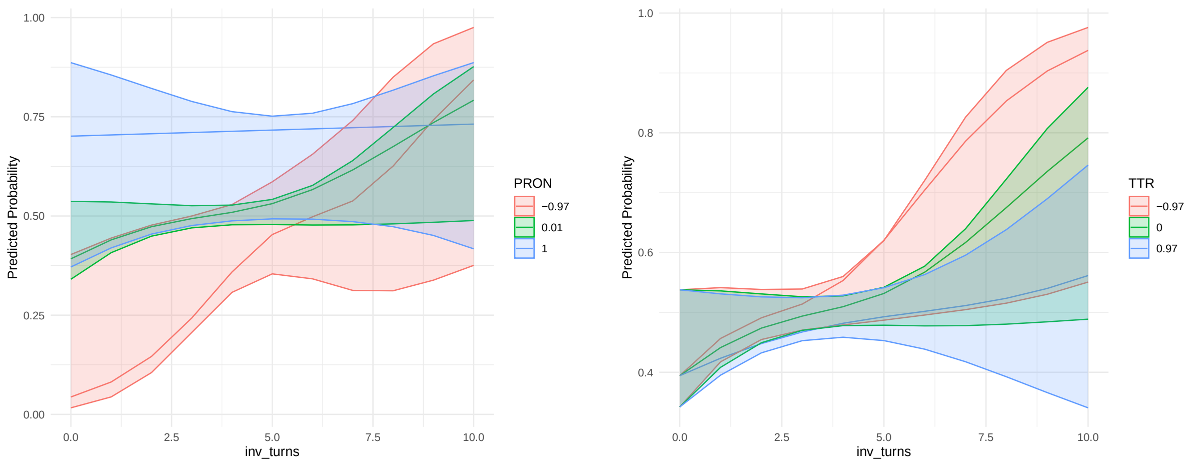
Table 2: The number of turns from participants and test administrators in the Pitt and the WLS corpus, before and after matching.

4.3 Quantifying the administrator effect

The Pitt model As shown in Figure 2(a), we found that the number of test administrators' turns remain positive and significant ($\beta = 0.331$, p-value < 0.05) in the $\mathcal{M}_{\text{pitt}}$, suggesting that a more interactive test administrator dynamic is associated with a higher probability of developing dementia. We also observed that pronoun usage ($\beta = 2.93$, p-value $<$ 0.001) showed a strong positive association with a higher probability of developing dementia. Interestingly, we observed significant interactions between test administrators' turns and various linguistic features, including TTR ($\beta = -0.146$, p-value $<$ 0.001), the usage of pronoun usage ($\beta = -0.330$, p-value < 0.05), auxiliary ($\beta = -0.417$, p-value < 0.001), adposition ($\beta = 0.382$, p-value < 0.05), and particle ($\beta = 0.191$, p-value < 0.001).

The WLS model As showed in Figure 2(b), we observed fewer significant predictors in $\mathcal{M}_{\text{wls}}$. Interestingly, we observed that, while the usage of pronoun ($\beta = -0.76$, p-value $<$0.1) showed significantly negative association with having a dementia diagnosis, its interactions terms with the number of test administrators' turns demonstrated an *opposite* directional effects ($\beta = 0.620$, p-value < 0.05).

The predicted effects of the interaction terms As shown in Figure 3(a), we observed that $\mathcal{M}_{\text{pitt}}$ predicts a dramatic increase in the probability of having a dementia diagnosis from 0.1 to 0.8 as conversations went longer for participants who used lower level of pronoun during the test. For participants with average pronoun usage (at mean, shown in green), $\mathcal{M}_{\text{pitt}}$ maintained consistent predicted probabilities of having a dementia diagnosis throughout all conversation lengths. Conversely, participants with high pronoun usage showed an initial high probability of approximately 0.8 for have a dementia diagnosis in shorter conversations, which gradually decreased to 0.7 as conversation went longer. As we observed in Figure 3(b), participants with lower TTR (shown in red) had an increasing probability of having a dementia diagnosis as the number of turns from test administrators increased, rising dramatically from around 0.5 to nearly 0.95 over 10 turns. Notably, participants with higher TTR (shown in blue) showed a contrasting pattern - their probability of having



(a) The effect of the interaction term between pronoun usage and `inv_turns`

(b) The effect of the interaction term between TTR and `inv_turns`

Figure 3: The predicted values and confidence intervals of the interaction terms between linguistic markers and `inv_turns`. The level of usage are denoted in color, where red indicates the lower usage (1 SD below the mean), green indicates the average usage, and blue indicates higher usage (1 SD above the mean). The x-axis indicates the number of turns from test administrators.

a dementia diagnosis actually decreased slightly as conversations went longer, dropping from 0.5 to 0.35. Furthermore, we found that the predictive probabilities of pronoun usage and TTR varies systematically with `inv_turns`. Collectively, Figure 3 suggests an interesting diagnostic transition: at a lower level of test administrator involvement (`inv_turns` ≤ 3, typical for healthy controls), pronoun usage provides greater diagnostic utility; at moderate involvement (`inv_turns` ≈ 4, typical for matched dementia patients), both features offer complementary values; while at a higher involvement levels ((`inv_turns` ≥ 6, typical for dementia patients before PSM), TTR becomes the dominant discriminative marker. This suggests that different linguistic features gain or lose diagnostic utility depending on the degree of administrator involvement.

4.4 Cross validation: classification performance

$\mathcal{M}_{\text{pitt}}$ achieved accuracy of 0.67, precision of 0.69, recall of 0.56, and F_1 score of 0.62 on the matched Pitt test split, respectively. Interestingly, $\mathcal{M}_{\text{pitt}}$ did not generalize well to the original WLS corpus, reaching accuracy of 0.59, precision of 0.13, recall of 0.38, and F_1 score of 0.19, respectively. $\mathcal{M}_{\text{pitt}}$ performed similarly on the matched WLS corpus, reaching accuracy of 0.50, precision of 0.50, recall of 0.38, and F_1 score of 0.43, respectively. $\mathcal{M}_{\text{wls}}$ also generalized poorly to the matched Pitt corpus, with accuracy of 0.55, precision of 0.54, recall of 0.47, and F_1 score of 0.50 on the matched Pitt test split.

5 Discussion

Our key findings are as follows. First, we show that many linguistic features previously studied in AD dementia progression appear to vary with level of test administrator involvement. Second, the observed variability between two corpora underscores the importance of considering administrator behavior as a potential confounding variable in linguistic analyses of clinical populations. These findings collectively suggest that some of the linguistic features commonly observed in dementia patients may be affected by the data collection processes rather than cognitive decline alone.

The observation of interactive test administrator dynamics in the Pitt corpus is consistent with prior work (Menn and Obler, 1989; Caplan and Hanna, 1998), which report that the test administrator needs to induce *enough* propositional utterances from participants in the constrained task such as the "Cookie Theft" picture description task. However, it is often noted that dementia patients are incapable of producing complex utterances due to the progression of the disease. As such, an interactive test administrator dynamic may lead to overestimation of a patients' linguistic ability in some cases.

Our results further suggest that these interaction patterns influence downstream dementia classification, which is consistent with a prior work (Farzana and Parde, 2022). Our study further quantifies the influence of test administrator behavior, demonstrating how the varying levels of investigator involvement between groups may confound our interpretation of linguistic markers as diagnostic indicators. Our results highlight the need to interpret linguistic markers not as isolated indicators, but as features embedded within an interactive context that includes test administrators' role in shaping the discourse. Further research design might benefit from explicitly accounting for and potentially controlling test administrator involvement when developing screening criteria based on linguistic features.

Our findings suggest a nuanced relationship between linguistic markers, administrator interaction patterns, and their predictability for cognitive decline. The consistently high predicted probability of a higher probability of developing dementia for participants with elevated pronoun usage (shown by the stable high probabilities in the blue line in Figure 3(a)) supports existing literature on pronoun over-usage (Almor et al., 1999; Jarrold et al., 2014; Cummings, 2019) as a linguistic marker of cognitive decline. However, our results also indicate that this relationship may be masked or amplified by test administrators' interaction styles, as evidenced by varying predicted probability trajectories across different conversation lengths. Similarly, while the observed TTR patterns also align with previous findings (Hier et al., 1985) that lower lexical diversity indicates cognitive decline, the dramatic increase in predicted probability for participants with lower TTR during longer conversations suggests that the established observations might be influenced by the test administrators' interaction patterns, suggests that these established linguistic markers may be partially attributable to differences in the test administration practices rather than the true construct measures of cognitive decline.

The disparities of classification performance of two models – $\mathcal{M}_{\text{pitt}}$ for *detecting* AD dementia, and $\mathcal{M}_{\text{wls}}$ for *predicting* dementia – confirms the often-observed challenges of developing robust and generalizable models for dementia detection and prediction. While $\mathcal{M}_{\text{pitt}}$ demonstrated moderate performance on its test split, it generalized poorly on the WLS corpus where precision and F_1 score dropped dramatically. $\mathcal{M}_{\text{pitt}}$'s slight improvement in performance on the matched vs. original WLS corpus suggests that the PSM may somewhat mitigate the confounding effect, but not fully resolve the cross-corpus and cross-task generalization issues. Similarly, $\mathcal{M}_{\text{wls}}$ showed limited generalization on the Pitt corpus. This consistent underperformance across corpora suggests the significant challenge of creating models that can *reliably* detect or predict dementia. Our results also suggest the need of considering corpus- and population-specific characteristics in the model development. Factors such as demographic differences, test administrating styles, and the temporal aspect of dementia progression (i.e., detection vs. prediction) may contribute to the observed lack of cross-corpus and cross-task generalizability.

The variability between two corpora suggests that some linguistic markers previously attributed to dementia may be specific to certain data collection protocols rather than universal linguistic anomalies associated with the disease's progression. $\mathcal{M}_{\text{pitt}}$ demonstrates reasonable performance on its own test split, suggesting that within a single dataset, certain linguistic patterns may indeed be indicative of cognitive decline after controlling for the influence of test administrators. However, its substantially degraded performance on the WLS corpus points out a critical issue: linguistic markers that appear robust within one population may not translate effectively to another. This lack of cross-corpus generalizability persists when we validate $\mathcal{M}_{\text{wls}}$ on the Pitt corpus - the performance of $\mathcal{M}_{\text{wls}}$ actually worsens on the matched Pitt test split. These findings collectively suggest that the linguistic anomalies associated with AD progression may be highly context-dependent, influenced by factors such as data collection protocols, test administrator dynamics, and population-specific characteristics. This indicates the need for caution when interpreting linguistic markers of cognitive decline, developing specialized neural language models, and validating findings across diverse datasets and populations.

While the speech samples produced by population with high clinical risks are scarce, incorporating text corpora drawn from different sources (also known as confounding by provenance) presents both opportunities and challenges for detecting linguistic anomalies in AD dementia. Previous studies demonstrate that treating the provenance of a transcript (i.e., Pitt vs. WLS) as a secondary target for prediction (Guo et al., 2021) and data augmentation (Liu et al., 2021; Bertini et al., 2022; Duan et al., 2023, *inter alia*) could lead to performance improvements. However, our results suggest the need for extra caution in such applications. These disparities suggest these approaches, if not carefully implemented, may introduce additional confounding variables rather than identifying true indicators of cognitive impairment. As such, the observed lack of cross-corpus and cross-task generalizability may explain why fine-tuned neural language models generalize less-than-ideal to other speech samples produced by populations at high clinical risk (Li et al., 2022; Farzana and Parde, 2023).

While the automated analysis of spoken language produced by population with high clinical risk remains a valuable component of early-screening cognitive assessment, the observed influence of test administrator dynamics on AD-related linguistic anomalies calls for a re-evaluation of current methods. Researchers and clinicians should exercise caution when interpreting the linguistic features of the "Cookie Theft" picture description task, as they may be partially artifacts of the data collection itself. Our results call for a standardized test administration to minimize the variability in administrator engagement, and the need for population- and language-specific norms for assessments.

6 Conclusion

Our study explored the relationship between test administrator involvement and linguistic features in dementia assessments using the "Cookie Theft" picture description task. The patterns we observed raise questions about how established linguistic features might be shaped by the dynamics of test administration alongside cognitive status. Our study brings the potential benefits of considering administrator behavior in future development of clinical speech analytics frameworks.

Limitations

The work presented here has several limitations. While our analysis identifies significant correlations between the test administrator interactions and linguistic features, we should note that our study design does not establish a direct causal link. Future experimental studies with standardized administrator protocols would be necessary to establish such a link. Second, the size of the datasets used in this study is considerably small, which is a common concern in this line of research (Petti et al., 2020). Moreover, all datasets used in this study are in American English, and many participants are representative of White, non-Hispanic American residents, which certainly limits the generalizability to other languages and ethnic groups. In this study, we only focus on analyzing POS tags for both datasets, which is a limited feature set for detecting cognitive impairment. Future studies should explore comprehensive linguistic and acoustic features (i.e., Fraser et al. (2015)) to establish a more definitive measurement of the effects of test administrator engagement. We acknowledge that there are linguistic differences between the two corpora studied in this work (Johnstone et al., 2015), which may affect the comparability of results across datasets. We should also note that while category fluency task has demonstrates the clinical utility for dementia screening; it is, however, not a complete clinical diagnosis, which may not capture the full spectrum of cognitive decline and could potentially lead to misclassification of some participants.

Acknowledgement

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01 LM014056-02S1, and the National Institute on Aging of the National institutes of Health under award number AG069792.

References

Hirotogu Akaike. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Amit Almor, Daniel Kempler, Maryellen C MacDonald, Elaine S Andersen, and Lorraine K Tyler. 1999. Why do alzheimer patients have difficulty with pronouns?

working memory, semantics, and reference in comprehension and production in alzheimer's disease. *Brain and language*, 67(3):202–227.

S. Ash, P. Moore, S. Antani, G. McCawley, M. Work, and M. Grossman. 2006. Trying to tell a tale. *Neurology*, 66(9):1405–1413.

Peter C. Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.

Silva Banovic, Lejla Junuzovic Zunic, and Osman Sinanovic. 2018. Communication difficulties as a result of dementia. *Materia socio-medica*, 30(3):221.

James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Archives of Neurology*, 51(6):585–594.

Flavio Bertini, Davide Allevi, Gianluca Lutero, Laura Calzà, and Danilo Montesi. 2022. An automatic alzheimer's disease classifier based on spontaneous spoken english. *Computer Speech & Language*, 72:101298.

Shauna Berube, Jodi Nonnemacher, Cornelia Demsky, Shenly Glenn, Sadhvi Saxena, Amy Wright, Donna C Tippett, and Argye E Hillis. 2019. Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American journal of speech-language pathology*, 28(1S):321–329.

Gerhard Blanken, Jürgen Dittmann, J-Christian Haas, and Claus-W Wallesch. 1987. Spontaneous speech in senile dementia and aphasia: Implications for a neurolinguistic model of language production. *Cognition*, 27(3):247–274.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Romola S Bucks, Sameer Singh, Joanne M Cuerden, and Gordon K Wilcock. 2000. Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Francisco Caamaño-Isorna, Montserrat Corral, Agustín Montes-Martínez, and Bahi Takkouche. 2006. Education and dementia: a meta-analytic study.

SJ Duff Canning, L Leach, D Stuss, L Ngo, and SE14981170 Black. 2004. Diagnostic utility of abbreviated fluency measures in alzheimer disease and vascular dementia. *Neurology*, 62(4):556–562.

David Caplan and Joy E. Hanna. 1998. Sentence production by aphasic patients in a constrained task. *Brain and Language*, 63(2):184–218.

Jane H Cerhan, Robert J Ivnik, Glenn E Smith, Eric C Tangalos, Ronald C Petersen, and Bradley F Boeve. 2002. Diagnostic utility of letter fluency, category fluency, and fluency difference scores in alzheimer's disease. *The Clinical Neuropsychologist*, 16(1):35–42.

Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and VJHW Welch. 2019. Cochrane handbook for systematic reviews of interventions. *Hoboken: Wiley*.

Catherine Crockford and Ruth Lesser. 1994. Assessing functional communication in aphasia: Clinical utility and time demands of three methods. *International Journal of Language & Communication Disorders*, 29(2):165–182.

Louise Cummings. 2019. Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia. *Pragmatics and Society*, 10(2):153–176.

Nilton Custodio, Lissette Duque, Rosa Montesinos, Carlos Alva-Diaz, Martin Mellado, and Andrea Slachevsky. 2020. Systematic review of the diagnostic validity of brief cognitive screenings for early dementia detection in spanish-speaking adults in latin america. *Frontiers in Aging Neuroscience*, 12.

Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. 2024. Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):1–43.

Junwen Duan, Fangyuan Wei, Jin Liu, Hongdong Li, Tianming Liu, and Jianxin Wang. 2023. CDA: A contrastive data augmentation method for Alzheimer's disease detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1819–1826, Toronto, Canada. Association for Computational Linguistics.

Eva Eggenberger, Katharina Heimerl, and Michael I. Bennett. 2012. Communication skills training in dementia care: a systematic review of effectiveness, training content, and didactic methods in different care settings. *International Psychogeriatrics*, 25:345 – 358.

Shahla Farzana and Natalie Parde. 2022. Are interaction patterns helpful for task-agnostic dementia detection? an empirical exploration. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 172–182, Edinburgh, UK. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2023. Towards domain-agnostic and domain-adaptive dementia detection from spoken language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11965–11978, Toronto, Canada. Association for Computational Linguistics.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2015. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, 49(2):407–422.

Letitia R Gewirth, Andrea G Shindler, and Daniel B Hier. 1984. Altered patterns of word associations in dementia and aphasia. *Brain and Language*, 21(2):307–317.

Harold Goodglass and Edith Kaplan. 1983. *Boston diagnostic aphasia examination booklet*. Lea & Febiger.

Melisa Gumus, Morgan Koo, Christa M. Studzinski, Aparna Bhan, Jessica Robin, and Sandra E. Black. 2024. Linguistic changes in neurodegenerative diseases relate to clinical symptoms. *Frontiers in Neurology*, 15.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "cookie theft" corpus chasm: applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3:642517.

Pamela Herd, Deborah Carr, and Carol Roan. 2014. Cohort Profile: Wisconsin longitudinal study (WLS). *International Journal of Epidemiology*, 43(1):34–41.

Daniel B Hier, Karen Hagenlocker, and Andrea Gellin Shindler. 1985. Language disintegration in dementia: Effects of etiology and severity. *Brain and language*, 25(1):117–133.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37, Baltimore, Maryland, USA. Association for Computational Linguistics.

Barbara Johnstone, Daniel Baumgardt, Maeve Eberhardt, and Scott Kiesling. 2015. *Pittsburgh speech and Pittsburghese*, volume 11. Walter de Gruyter GmbH & Co KG.

William Labov. 1973. *Sociolinguistic patterns*. 4. University of Pennsylvania press.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Changye Li, Weizhe Xu, Trevor Cohen, Martin Michalowski, and Serguei Pakhomov. 2023. Trestle: Toolkit for reproducible execution of speech, text and language experiments. *AMIA Summits on Translational Science Proceedings*, 2023:360.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. 2021. Detecting alzheimer's disease from speech using neural networks with bottleneck features and data augmentation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7323–7327.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Proc. Interspeech 2020*, pages 2172–2176.

Lise Menn and Loraine K Obler. 1989. Cross-language data and theories of agrammatism. In *Agrammatic aphasia*, pages 1369–1389. John Benjamins.

Andreas U Monsch, Mark W Bondi, Nelson Butters, David P Salmon, Robert Katzman, and Leon J Thal. 1992. Comparisons of verbal fluency tasks in the detection of dementia of the alzheimer type. *Archives of neurology*, 49(12):1253–1258.

Tiia Ngandu, Eva von Strauss, E-L Helkala, B Winblad, A Nissinen, J Tuomilehto, H Soininen, and M Kivipelto. 2007. Education and dementia: what lies behind the association? *Neurology*, 69(14):1442–1450.

Thu T Nguyen, Eric J Tchetgen Tchetgen, Ichiro Kawachi, Stephen E Gilman, Stefan Walter, Sze Y Liu, Jennifer J Manly, and M Maria Glymour. 2016. Instrumental variable approaches to identifying the causal effect of educational attainment on dementia risk. *Annals of epidemiology*, 26(1):71–76.

Marjorie Nicholas, Loraine K. Obler, Martin L. Albert, and nancy helm estabrooks. 1985. Empty speech in alzheimer's disease and fluent aphasia. *Journal of speech and hearing research*, 28 3:405–10.

Serguei Pakhomov, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. 2011. Computerized assessment of syntactic complexity in alzheimer's disease: a case study of iris murdoch's writing. *Behavior research methods*, 43:136–144.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

Ulla Petti, Simon Baker, Anna Korhonen, and Jessica Robin. 2023. How Much Speech Data Is Needed for Tracking Language Change in Alzheimer's Disease? A Comparison of Random Length, 5-Min, and 1-Min Spontaneous Speech Samples. *Digital Biomarkers*, 7(1):157–166.

Sandeep Reddy. 2022. Explainability and artificial intelligence in medicine. *The Lancet Digital Health*, 4(4):e214–e215.

Marc Rousseaux, Amandine Sève, Marion Vallet, Florence Pasquier, and Marie Anne Mackowiak-Cordoliani. 2010a. An analysis of communication in conversation in patients with dementia. *Neuropsychologia*, 48(13):3884–3890.

Marc Rousseaux, Amandine Sève, Marion Vallet, Florence Pasquier, and Marie Anne Mackowiak-Cordoliani. 2010b. An analysis of communication in conversation in patients with dementia. *Neuropsychologia*, 48(13):3884–3890.

Annemieke Ruitenberg, Alewijn Ott, John C. van Swieten, Albert Hofman, and Monique M.B. Breteler. 2001. Incidence of dementia: does gender make a difference? *Neurobiology of Aging*, 22(4):575–580.

Steven R Sabat. 1994. Language function in alzheimer's disease: a critical review of selected literature. *Language & Communication*, 14(4):331–351.

Michal Tomek Seyed Ahmad Sajjadi, Karalyn Patterson and Peter J. Nestor. 2012. Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, 26(6):847–866.

Mengke Shi, Gary Cheung, and Seyed Reza Shahamiri. 2023. Speech and language processing with deep learning for dementia diagnosis: A systematic review. *Psychiatry Research*, 329:115538.

David A Snowdon, Susan J Kemper, James A Mortimer, Lydia H Greiner, David R Wekstein, and William R Markesbery. 1996. Linguistic ability in early life and cognitive function and alzheimer's disease in late life: Findings from the nun study. *Jama*, 275(7):528–532.

Laura Stokes, Helen Combes, and Graham Stokes. 2015. The dementia diagnosis: a literature review of information, understanding, and attributions. *Psychogeriatrics*, 15(3):218–225.

Wiesje M van der Flier and Philip Scheltens. 2005. Epidemiology and risk factors of dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 5):v2–v7.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jochen Weiner, Christian Herff, and Tanja Schultz. 2016. Speech-based detection of alzheimer's disease in conversational german. In *Interspeech 2016*, pages 1938–1942.

Zhongheng Zhang, Hwa Jung Kim, Guillaume Lonjon, Yibing Zhu, et al. 2019. Balance diagnostics after propensity score matching. *Annals of translational medicine*, 7(1).

Appendix

POS tags	**Name**
ADJ	Adjective
ADP	Adposition
ADV	Adverb
AUX	Auxiliary
CCONJ	Coordinating conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
PART	Particle
PRON	Pronoun
PROPN	Proper noun
SCONJ	subordinating conjection
VERB	Verb

Table 3: The Universal POS tags

Features	*Before matching*				*After matching*			
	Level	Control	Dementia	SMD	Level	Control	Dementia	SMD
Number of transcripts (n)		182	214			167	167	
Education (mean (SD))		13.92 (2.42)	12.28 (2.81)	0.629		13.66 (2.24)	12.53 (2.93)	0.434
Age (mean (SD))		64.08 (7.91)	71.51 (8.63)	0.897		64.27 (7.85)	71.46 (8.63)	0.871
Gender (%)	Female	114 (62.6)	147 (68.7)	0.128	Female	104 (62.3)	116 (69.5)	0.152
	Male	68 (37.4)	67 (31.3)		Male	63 (37.7)	51 (30.5)	
PRON (mean (SD))		15.03 (9.85)	17.18 (12.36)	0.193		14.72 (9.48)	15.59 (10.80)	0.086
PROPN (mean (SD))		0.12 (0.51)	0.25 (0.65)	0.227		0.13 (0.53)	0.14 (0.46)	0.024
NOUN (mean (SD))		24.93 (13.94)	19.41 (11.06)	0.439		24.57 (13.96)	19.37 (10.88)	0.416
ADJ (mean (SD))		4.06 (3.52)	3.21 (3.47)	0.243		3.88 (3.38)	3.16 (3.48)	0.211
ADV (mean (SD))		3.91 (3.75)	5.43 (5.03)	0.342		3.81 (3.79)	4.65 (3.95)	0.217
CLAUSE (mean (SD))		20.13 (9.22)	20.43 (11.00)	0.030		19.72 (8.92)	18.87 (9.21)	0.093
AUX (mean (SD))		13.18 (6.36)	11.66 (7.09)	0.224		13.02 (6.32)	11.22 (6.55)	0.281
VERB (mean (SD))		16.81 (8.17)	15.70 (8.79)	0.131		16.49 (7.94)	15.00 (8.06)	0.186
ADP (mean (SD))		11.58 (7.22)	9.29 (6.32)	0.338		11.35 (7.05)	9.50 (6.40)	0.274
DET (mean (SD))		16.65 (9.07)	13.73 (7.97)	0.342		16.40 (9.01)	13.79 (8.05)	0.306
PUNCT (mean (SD))		24.41 (10.80)	23.96 (12.13)	0.040		24.23 (10.67)	22.23 (9.90)	0.195
CCONJ (mean (SD))		5.68 (4.28)	5.84 (4.15)	0.038		5.59 (4.20)	5.85 (4.21)	0.063
PART (mean (SD))		2.77 (2.25)	3.21 (2.74)	0.174		2.59 (2.11)	3.09 (2.50)	0.214
SCONJ (mean (SD))		1.63 (2.46)	1.27 (1.78)	0.171		1.58 (2.46)	1.18 (1.72)	0.189
INTJ (mean (SD))		5.16 (4.02)	6.21 (6.83)	0.187		5.07 (3.97)	5.66 (4.61)	0.138
LF (mean (SD))		8.16 (0.36)	8.36 (0.47)	0.479		8.15 (0.37)	8.30 (0.45)	0.358
TTR (mean (SD))		0.33 (0.05)	0.31 (0.06)	0.373		0.34 (0.05)	0.32 (0.06)	0.286
par_turns (mean (SD))		13.55 (6.04)	13.54 (6.98)	0.003		13.44 (5.97)	12.38 (5.60)	0.183
inv_turns (mean (SD))		3.16 (1.77)	6.10 (4.48)	0.863		3.33 (1.73)	4.38 (1.85)	0.589
mmse (mean (SD))		29.13 (1.11)	18.54 (5.11)	2.864		29.08 (1.13)	19.50 (4.50)	2.920

Table 4: The differences of linguistic features before/after matching on the Pitt corpus

Features	*Before matching*				*After matching*			
	Level	Control	Dementia	SMD	Level	Control	Dementia	SMD
Number of transcripts (n)		1017	152			152	152	
Education (mean (SD))		13.77 (3.01)	12.64 (2.16)	0.431		12.62 (2.18)	12.64 (2.16)	0.006
Age (mean (SD))		70.30 (4.14)	70.20 (5.75)	0.021		70.81 (3.77)	70.20 (5.75)	0.126
PRON (mean (SD))		15.20 (9.93)	11.16 (8.05)	0.447		14.37 (8.91)	11.16 (8.05)	0.377
AUX (mean (SD))		10.76 (6.63)	7.81 (5.48)	0.485		9.75 (6.25)	7.81 (5.48)	0.330
VERB (mean (SD))		16.82 (9.08)	12.71 (7.19)	0.502		15.25 (7.91)	12.71 (7.19)	0.336
ADP (mean (SD))		11.53 (6.74)	8.57 (5.71)	0.474		10.07 (5.77)	8.57 (5.71)	0.262
DET (mean (SD))		16.99 (9.87)	12.22 (7.23)	0.551		15.03 (8.00)	12.22 (7.23)	0.368
NOUN (mean (SD))		29.00 (16.89)	22.38 (13.71)	0.430		26.86 (14.61)	22.38 (13.71)	0.316
PUNCT (mean (SD))		26.61 (13.24)	21.68 (11.74)	0.393		25.47 (11.87)	21.68 (11.74)	0.321
CCONJ (mean (SD))		5.47 (4.83)	3.43 (3.62)	0.478		5.02 (4.99)	3.43 (3.62)	0.365
ADJ (mean (SD))		3.89 (3.75)	2.30 (2.37)	0.509		3.02 (2.93)	2.30 (2.37)	0.272
PART (mean (SD))		2.83 (2.46)	2.40 (2.14)	0.186		2.63 (2.25)	2.40 (2.14)	0.105
SCONJ (mean (SD))		1.55 (1.92)	0.90 (1.36)	0.390		1.38 (1.60)	0.90 (1.36)	0.324
ADV (mean (SD))		4.06 (3.98)	2.93 (3.41)	0.305		3.77 (3.60)	2.93 (3.41)	0.240
INTJ (mean (SD))		1.88 (2.91)	1.50 (2.58)	0.139		1.99 (2.80)	1.50 (2.58)	0.181
LF (mean (SD))		8.06 (0.43)	8.05 (0.44)	0.033		8.02 (0.42)	8.05 (0.44)	0.057
TTR (mean (SD))		0.35 (0.06)	0.37 (0.07)	0.303		0.36 (0.06)	0.37 (0.07)	0.189
CLAUSE (mean (SD))		21.01 (10.44)	17.21 (9.26)	0.385		19.81 (9.14)	17.21 (9.26)	0.282
PROPN (mean (SD))		0.07 (0.36)	0.02 (0.18)	0.170		0.14 (0.55)	0.02 (0.18)	0.288
par_turns (mean (SD))		14.39 (7.91)	11.97 (7.04)	0.323		13.88 (6.68)	11.97 (7.04)	0.278
inv_turns (mean (SD))		0.75 (1.53)	0.82 (1.79)	0.044		0.77 (1.25)	0.82 (1.79)	0.034

Table 5: The differences of linguistic features before/after matching on the WLS dataset