LINGUISTIC FEATURES IN MILD AND MODERATE DEMENTIA

Alexsandra Murphy

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation
Murphy, Alexsandra, "LINGUISTIC FEATURES IN MILD AND MODERATE DEMENTIA" (2024). Electronic Theses and Dissertations. 3598.
https://digitalcommons.memphis.edu/etd/3598
LINGUISTIC FEATURES IN MILD AND MODERATE DEMENTIA

by

Alexandra Murphy

A Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

Major: Applied Linguistics

The University of Memphis
August 2024
Preface

This paper was completed in partial fulfillment for the Master of Arts degree in Applied Linguistics at the University of Memphis. This paper examines the linguistic variations in a healthy control group and a clinical group with mild and moderate dementia. Although this study uses a replicated methodology, we have the hope that more studies using a corpus-based and computational linguistic approach will create a more robust body of research and that such methodologies can eventually predict the onset of Alzheimer’s dementia.
Abstract

The aim of this study is to understand what linguistic measures differ in individuals who have dementia versus a healthy control group as linguistic measures can provide insight to discrepancies in cognitive abilities, notably language. The problem with identifying and diagnosing diseases related to memory, such as dementia and Alzheimer’s disease, is that diagnoses require involved medical examinations, such as neuroimaging, and descriptive diagnostics that may be subjective and relative. This can lead to ambiguous results and lack of clarity for patients and their families and medical care team, making it difficult to chart a course of action or make decisions in the patient’s best interest. However, if we take a methodical, quantitative linguistic look at the language used in standard memory care diagnoses, we can generate objective parameters or baselines for rates of word usage that can provide more clarity for patients and their families and medical care teams. By using language categories and baseline rates of usage of terms like articles, prepositions, and dysfluencies in control populations versus those individuals with definitive mild and moderate memory care diagnoses, we can provide clinicians and practitioners with guidelines for making more objective diagnoses. Alongside other diagnostic processes, linguistic analysis will help the medical care team better inform patients and their families so they can make better choices and decisions about the level of care and interventions they need. In this study we analyze the lexical categories of patients diagnosed with mild and moderate Alzheimer’s/dementia (hereafter AD) alongside a control group to uncover patterns of language usage. We use a computational, corpus-based approach to gauge which psycholinguistic categories and subcategories and linguistic features are reflected in the speech of patients with mild and moderate dementia, with the help of the LIWC (Linguistic Inquiry and Word Count) software.
## Table of Contents

List of figures and tables  iv

**Introduction**  pg. 1  
**Literature Review**  pg. 2  
*Alzheimer’s and Dementia: The disease*  pg. 2  
*Normal versus disordered cognitive decline*  pg. 3  
*Mini Mental State Examination*  pg. 5  
*Language and Dementia*  pg. 6  
*Corpus-based approaches to AD research*  pg. 9  
**Data**  pg. 15  
**Methods**  pg. 16  
*The Cookie Theft Picture*  pg. 16  
*Linguistic Inquiry and Word Count*  pg. 19  
**Results and Discussion**  pg. 20  
**Conclusion and Future directions**  pg. 27  

**References**  pg. 29  
**Appendix 1**  pg. 32  
Categories and subcategories of LIWC dimensions
List of Figures and Tables

Figure 1. The Cookie Theft Picture
Figure 2. Control group versus Mild AD group
Figure 3. Control group versus moderate AD group
Figure 4. Mild AD group versus moderate AD
Figure 5. Language trends across all 3 participant groups

Table 1. Categories and subcategories of LIWC dimensions
Introduction

Dementia is a general term used for the impairment to remember, think, and make decisions. Older populations are more widely affected by dementia, but it is important to note that Dementia is different from normal, cognitive decline that occurs solely due to aging. According to the Centers for Disease Control and Prevention (2023), the number of adults (at least 65 years of age) affected by Alzheimer’s dementia is projected to be close to 14 million by the year 2060.

Common symptoms of dementia include issues with memory, attention, communication, reasoning, judgment, problem solving, and visual perception (Centers for Disease Control and Prevention, 2023). Risk factors include age, family history, race and ethnicity, poor heart health, and traumatic brain injury. Treatment of dementia depends on the underlying cause of its development.

Neurologically, Alzheimer’s dementia causes widespread damage in the brain, resulting in poor neuron functioning. Alzheimer’s dementia tends to damage the connections among neurons that relate to brain areas involved in memory as well as language and social behaviors (National Institute of Aging, What Happens to the Brain in Alzheimer's Disease?) For diagnosis of dementia, diagnostic criteria include memory impairment as well as impairment in at least one other domain of cognitive function, and may include: “aphasia, apraxia, agnosia, or another sort of disturbance to executive functioning” (Wetherell and Jeste, 2022). The present deficits must present a decline compared to a previous level of functioning and the level of severity must cause significant impairments in either social or occupational performance. Generally, the diagnosis of dementia begins with a person presenting memory difficulties alongside other issues that may include: “apathy, lack of initiative, disorientation, sleeping disturbances, aggression, depression
and anxiety or other psychotic symptoms, attention and concentration, language, motor coordination, visuospatial skills, and insight, and judgment” (Wetherell, 2003).

The current study uses a corpus-based approach with the software Linguistic Inquiry and Word Count (LIWC), a text analysis program, to examine any differences present in the speech of 3 groups: a healthy control group, a group with mild dementia, and a group with moderate dementia. We hypothesize that participants with mild and moderate dementia will use more “filler words” in their speech than participants in the healthy control group. We predict that participants with mild and moderate dementia will use more “nonfluencies” and fewer words overall. Essentially, we predict that word retrieval will be more difficult for the groups with mild and moderate AD than the healthy, control group. The objective of this paper is to answer the question: do individuals with mild and moderate dementia use significantly different descriptors when describing a visual image than a healthy control group?

This paper proceeds as follows: We first provide a brief review of literature regarding Alzheimer’s dementia and then a review of literature of projects with similar datasets and methods to address linguistic variations in AD. We also provide a description of the current study’s dataset and methodology, then provide results and a discussion. We conclude with a brief conclusion covering limitations and future directions.

**Literature Review**

**Alzheimer’s and Dementia: The disease**

The Alzheimer’s Association classifies the disease into three categories: Mild, moderate, and severe (Stages of Alzheimer’s, Alzheimer’s Association, 2024). During the first stage, symptoms might not be noticeable, and the person may still be able to function independently. At this point, the affected individual may have difficulty retrieving a particular word, be unable to
remember names upon meeting new people, have difficulties with tasks in social or work settings, misplace objects, or experience trouble with planning or organizing.

The middle stage of Alzheimer’s (moderate) usually lasts the longest out of the stages and may even persist for many years. As the disease progresses, the individual loses their independence and needs more in-depth care. At this point in the illness, individuals may be forgetful of their own personal history, have emotional or behavioral issues, be unable to dress appropriately for the weather/climate, lose control over bodily functions, or become lost easily in familiar settings.

Late stage (severe) Alzheimer’s is the final stage of the illness. During this stage, people with Alzheimer’s dementia may require full-time care. They may not be able to recount recent experiences or recognize their surroundings. Individuals may also experience difficulties in their physical abilities, such as walking, sitting, and even swallowing. Individuals also experience issues in communicating and may have a more vulnerable immune system. During the final stage, palliative care for the patient eventually becomes necessary (Stages of Alzheimer’s, Alzheimer’s Association, 2024).

Normal versus disordered cognitive decline

It is important to note that cognitive decline is, to a degree, a normal part of healthy aging. However, there are stark differences that separate normal aging and dementia disorders. Alzheimer’s disease has become a global health issue as more of the population ages. There is no effective treatment for Alzheimer’s disease to stop progression of symptoms (Lo, 2017). Some researchers believe that the lack of treatment is a result of the delays in diagnosis of the illness. Generally, by the time an individual is diagnosed with Alzheimer’s the neurodegenerative changes are at a stage where few neurons can be salvaged with medication.
The Alzheimer’s research community has introduced the concept of Mild Cognitive Impairment (MCI) to identify pre-dementia patients who could potentially benefit from therapeutic drugs. MCI is heterogeneous in terms of its underlying pathology and its practicality in predicting dementia. Lo (2017) introduces the idea of MCI as being a “borderland” between healthy aging and dementia. Essentially, MCI is a concept that has been used in an attempt to identify patients with Alzheimer’s early on in the course of the disease. The diagnosis of MCI versus dementia depends not only on cognitive tests, but also on the interaction of individuals and their sociocultural environment. The amount of cognitive exertion is contingent on one’s demands in their environment. Some cognitive changes during aging are normal and not a cause for concern.

Normal age-related changes usually pertain to thinking speed and attention. Mild cognitive impairment refers to decline that does not interfere with a someone’s ability to complete typical, everyday tasks. In abnormal aging, changes are more rapid and more severe. Abnormal aging can also include motor issues and people experiencing abnormal cognitive decline may have issues tripping or falling frequently. Symptoms seemingly increase rapidly in the years after diagnosis. Certain factors may contribute to cognitive decline such as having type 2 diabetes, high blood pressure, midlife obesity, smoking, depression, little or no mental stimulation, and little or no physical activity (UCSF, Weill Institute for Neurosciences).

Generally, cognitive, functional, and behavioral tests to assess dementia include: Ascertain Dementia 8 (AD8) Functional Activities Questionnaire (FAQ), Mini-Cog, Mini Mental-State Exam (MMSE), Montreal Cognitive Assessment (MoCA), and Neuropsychiatric Inventory Questionnaire (NPI-Q). Structural imaging is also commonly used in diagnosis of dementia with the main purpose being to rule out any other diseases that may have similar
symptoms to dementia. For example, magnetic resonance imaging (MRI) and computed
tomography (CT) can reveal tumors, evidence of prior stroke, or damage from head trauma
(Alzheimer’s association).

Over the last decade, computational methodologies have grown in popularity as a means
to detect linguistic differences in Alzheimer’s Dementia (Eyigoz et al., 2020, Orimaye et al.,
2017, and Kempler and Goral, 2008). Since dementia affects language abilities, there is research
that assesses language specifically as opposed to using standard diagnostic criteria and
neuropsychological examinations (Orimaye et al., 2017).

*Mini Mental State Examination*

The Mini-Mental State Examination (MMSE) was created by Marshal Folstein et al. in
1975 as an instrument for brief assessment of mental status in hospitalized patients. It is the most
widely used test for standardized cognitive assessment in clinical settings, particularly within the
elderly population. It has been translated into multiple languages. It is ultimately widely used in
the scientific community (Gallegos et al., 2022).

The Mini-Mental State Examination (MMSE) is the best known and most often utilized
screening tool for providing an overall measure of cognitive impairment in clinical, research, and
community settings (Arevalo-Rodriguez et al., 2021). The MMSE is used to screen patients for
cognitive impairment, track changes in cognitive functioning over time, and to often assess the
effects of therapeutics agents on cognitive functioning. Since the development of the MMSE,
there has been a plethora of literature published on the MMSE and it has been demonstrated to
be a relatively sensitive marker of overt dementia (O’Bryant et al., 2008).

However, its utility decreases when used to assess patients with mild cognitive decline
and other psychiatric conditions (O’Bryant et al., 2008). Although the examination has been
modified to a degree over time, the test has always consisted of 2 parts. The first part of the examination asks questions related to orientation, memory, and attention. The second portion of the examination assesses verbal and written ability of the examinee (Gallegos et al., 2022). A score of 30 indicates that an individual does not have current cognitive decline. A score of 26 to 29 may be deemed “questionable,” 21 to 25 indicates a mild score of dementia, 11 to 20 indicates moderate dementia, and a score of 0 to 10 indicates severe dementia (Perneczky et al., 2006).

Language and Dementia

A hallmark characteristic of Dementia is cognitive decline which largely entails language deficits. Common language deficits in dementia include problems with word retrieval, deficits in sentence comprehension, and lack of cohesion in discourse (Kempler and Goral, 2008). These language deficits in dementia occur in the context of multiple cognitive impairments. Kempler and Goral (2008) review findings of language impairment in three different dementia syndromes. Within each dementia type, the affected areas of the brain vary.

In dementia of the Alzheimer’s type, there is a progressive deterioration of memory with at least two other cognitive domains, such as language, visuospatial perception, or executive function. The hippocampus and frontal cortex are affected in Alzheimer’s disease. Frontotemporal dementia has two varieties, and they are distinguished from other types of dementia syndromes due to their marked language impairments. One type of frontotemporal dementia is semantic dementia and is characterized by fluent speech output accompanied by anomia and comprehension impairments.

The other frontotemporal variant, progressive nonfluent aphasia, is characterized by nonfluent speech output and anomia alongside relatively preserved comprehension. Kempler and
Goral (2008) discuss the lexical, sentence, and discourse impairments of each syndrome. They note that the data suggest that these three different syndromes vary regarding the contribution of grammatical, semantic, and conceptual, and extralinguistic deficits to the observed language impairments. Overall, the data suggest that many language impairments within dementia are due to extralinguistic factors, such as issues with episodic memory.

Issues with memory and attention impact word finding in early and moderate dementia Alzheimer's type. Decreases in executive function and memory cause sentence-level processing issues that are found in all three syndromes. Deficits in the semantic/conceptual system itself are predominantly responsible for the naming and word comprehension impairments in semantic dementia and later stage dementia Alzheimer's type. The close relationship among executive function, language abilities, and bilingualism all support the concept that cognitive reserve can delay the onset of dementia however, it is unclear how cognitive reserve is acquired and the role of such factors as genetic predispositions, formal education, and sociolinguistic environment (Kempler and Goral, 2008).

Emery (2000) completed a meta-analysis of AD research and applied a semiotic language framework to organize findings on language impairment in Dementia Alzheimer's Type (DAT). Memory impairment is widely known to be the most glaring symptom of Dementia. However, it is now known that language deficits occur in around 8 to 10% of Alzheimer’s patients as a primary symptom in the earlier stages of the disease. Emery (2000) examines the nature of language impairment in DAT within the context of an ordered language hierarchy. Relationships between the differing kinds of language deficits in dementia can be better understood within a hierarchical context. Emery (2000) describes language as a system of codes evolved for the purpose of communication. It is a cognitive activity that allows people to speak utterances that
have neither been heard nor rehearsed, which still conforms to a set of rules. Emery (2000) analyzed research studies that spanned over the course of 40 years. Emery (2000) explains that research papers were chosen for review if references were made to any or more of the language parameters of: phonology, morphology, syntax, semantics, as well as sound production, naming, grammar, sentence processing, and verbal comprehension in Alzheimer’s patients.

One of Emery’s (2000) main take-aways from the meta-analysis, is that phonological impairment in Alzheimer’s is under-documented and can affect patients’ abilities to process, reproduce, or use normal speech sounds. Most of the research related to phonological impairments suggests that Alzheimer’s patients retain the ability to produce the sounds of their spoken language appropriately until advanced stages in the illness. However, Emery (2000) explains that it has been noted that Alzheimer’s patients sometimes have the preserved ability to continue to read aloud even if they have no understanding of the words they are reading. In morphologic impairments in dementia, complexity tends to be an intervening variable between task demands and good performance on the task.

It seems that the more complicated a task is, the worse the task performance will be. Alzheimer’s patients seem to do reliably worse at the morphologic rank of the language hierarchy than at the less complex rank of phonology. Regarding syntax, Emery explains that there is no question that Alzheimer’s patients can generally produce syntactically acceptable vocalizations, especially ones that are simpler and automatic if they are able to use language still. It also has been suggested that syntactic production in patients with Alzheimer’s consist mostly of programmed utterances. Regarding semantics in Alzheimer’s patients, it has been noted that the best Alzheimer’s patients’ performance at the level of vocabulary occurs under the conditions of automatized use of lexicon in which integration between meaning and sound should be
minimal. With larger aggregates of words, data suggests that the more abstract, logical, and complex a sentence is, the more issues that an Alzheimer’s patient would have deciphering it. Maximal semantic processing seems to be related to simplicity and concreteness. Emery (2000) states there appears to be an overall relationship between language decline and the complexity of language across the hierarchical ranks. There also seems to be a negative relationship between the sequence in language development and language decline. Language forms that are learned last in the sequence of language development are the most complex and appear to be the first to deteriorate alongside cognitive decline in Alzheimer’s (Emery, 2000).

*Corpus-based approaches to AD research*

Ferguson et al., (2009) discuss the importance of using corpus-based research methods for research within speech-pathology and similar research fields. They note that people have different patterns within their speech which may be contingent on varying levels of proficiency that have been acquired through childhood as well as second language acquisition. Changes also occur alongside healthy aging and abnormal changes such as dementia or other acquired pathologies such as brain injuries.

Corpus-based linguistic approaches have provided complementary methodologies to examining language variation. In the field of speech-language pathology, corpus-based research relies heavily on archival databases. Computer software is widely used to analyze corpora since corpora generally consist of a large amount of data and, analyzing such data manually would be quite labor-intensive. Corpus linguistics has been applied to fields such as teaching English for academic purposes and teaching English as a second language. Fields such as these have established a reference point for what is a “normal” or an “ideal” target level of performance in order to see where an individual diverges from that reference point. They note that corpus-based
research methods in linguistics have been adopted in clinical settings by speech-language pathologists.

The Brown Corpus of American English was used for developing stimuli for assessments of naming in aphasia. Word finding difficulty, or anomia, is a common issue in aphasia and is known to be affected by word frequency. Data from Francis and Kucera was used to develop lists of high and low frequency words within the test batteries for the Psycholinguistic Assessments of Language Processing in Aphasia. In the field of speech-language pathology, the Child Language Data Exchange System (CHILDES) database has been widely used and has been incorporated into the larger database TalkBank. TalkBank is an interdisciplinary research database that was developed by Dr. Brian MacWhinney. The goal of TalkBank is to cultivate fundamental research in the study of human communication. Since the inception of the CHILDES database, its system has enabled the development of international standards for transcription and analysis: Codes for the Human Analysis of Transcripts (CHAT) and Computerized Language Analysis (CLAN) (Ferguson et al., 2009).

Eyigoz et al., (2020) used a corpus method alongside other linguistic classification methods to predict the future onset of Alzheimer’s disease in cognitively normal participants. They used predictive modeling for a possible, future, diagnosis of Alzheimer’s. To study linguistic performance as an early biomarker of Alzheimer’s, they used predictive modeling of future diagnosis of Alzheimer’s from a cognitively normal baseline of Framingham Heart Study participants. The linguistic variables that they used were derived from written responses to “The Cookie-Theft” picture description task. They compared the predictive performance of linguistic variables with clinical and neuro-physical variables. Their sample size consisted of 703 samples from 270 participants out of which a dataset consisting of a single sample from 80 participants
was held out for testing. Half of the participants in the test set developed Alzheimer’s symptoms before 85 years of age while the other half did not. Eyigoz et al., (2020) collected all of their data during the sample set’s “cognitively normal” period. They note that a priority in Alzheimer’s Disease research is the identification of early intervention strategies that will ultimately decrease the risk, delay the onset, or slow the progression of the disease. They note that the Framingham Heart Study is a well-documented and community-based, cohort study initiated in 1948 with the purpose of monitoring participants' health.

To examine psycholinguistic factors, Eyigoz et al., (2020) examined verbosity, lexical richness, and repetitiveness in patients with cognitive decline. This was measured by using metrics such as number of words, number of unique words, and frequencies of repetition. Their results suggest that language performance in naturalistic probes expose subtle early signs of progression to Alzheimer’s disease in advance of clinical diagnosis of impairment (Eyigoz, et al., 2020).

Orimaye et al., (2017) used a corpus-based approach to examine predictability of Alzheimer’s dementia by focusing on linguistic deficits and biomarkers. They explain that the diagnosis of neurodegenerative disorders, such as Alzheimer’s disease and related dementias, has posed challenges. Neurodegenerative disorders are typically diagnosed using specific clinical diagnostic criteria and neuropsychological examinations. However, the use of several Machine Learning Algorithms to build automated diagnostic models using low-level linguistic features that result from verbal utterances could aid diagnosis of individuals with probable Alzheimer’s.

Orimaye et al., (2017) developed different machine learning models on the Dementia TalkBank language transcript clinical dataset which consisted of 99 patients with probably Alzheimer’s and 99 healthy controls. Their models learned several syntactical, lexical, and n-
gram linguistic biomarkers to differentiate between the probable Alzheimer’s group and the control group. Their models learned several syntactic, lexical, and n-gram linguistic biomarkers in order to distinguish the probable Alzheimer’s diseases group from the healthy group.

Orimaye et al., (2017) found that the group with probable Alzheimer’s had significantly less usage of syntactic components and significantly higher usage of lexical components in their language. There was a significant difference in the use of n-grams as the group of healthy individuals were able to identify and make sense of more objects in their n-grams than the probable Alzheimer’s group. Their best diagnostic model significantly distinguished the probable Alzheimer’s group from the healthy group with a better Area Under Receiving Operating Characteristics Curve (AUC) using Support Vector Machines. They conclude that experimental and statistical evaluations suggest that using machine learning algorithms for learning linguistic biomarkers from the verbal utterances of elderly people may assist in the clinical diagnosis of probable Alzheimer’s disease (Orimaye et al., 2017).

Shibata et al., (2016) also used a corpus-based approach and a computational analysis to examine differences between a group with Alzheimer’s dementia and a healthy control group. Shibata et al., (2016) explain that detecting Alzheimer’s disease in early stages based on natural language processing (NLP) has drawn attention in recent years. Vocabulary size, grammatical complexity and fluency have all been examined through Natural Language Processing (NLP) metrics. They note that the content analysis of Alzheimer’s disease patients in used in their spoken language narratives in still unreachable in terms of NLP. The study investigated features of language used by Alzheimer’s patients. The sample size consisted of 18 examinees, ranging between the ages of 53 to 90. Individuals were separated into two groups: the Alzheimer’s group consisted of 9 individuals with Mini Mental State Examination (MMSE) scores of 21 or lower.
The healthy control group consisted of 9 participants with MMSE scores of 22 or higher. They use Linguistic Inquiry and Word Count (LIWC) to analyze narratives of people who are suspected to have Alzheimer’s disease. They collected narratives from hospital patients in order to build a corpus. Patients’ conversations with medical staff were recorded and then transcribed.

Shibata et al., (2016) found that values in the “social” LIWC category were used significantly less by Alzheimer’s patients than participants in the healthy control group. They mention that there are correlations between little social contact and the development of Alzheimer’s. They state that their result aligns with the idea that little social contact may lead to Alzheimer’s. They also found that “Ipron” (impersonal pronouns) were significantly higher in the Alzheimer’s group than in the healthy control group. They suggest that this is because individuals with Alzheimer’s are forgetful of words and may use impersonal pronouns to supplement for other nouns. The value of the category’s verbs and elements of present tense (“present focus”) were significantly larger than those in the healthy control. They describe their report as the first of a quantitative study that investigated the word categories of Alzheimer’s disease. Significant differences were found for the Alzheimer’s group in the usage of multiple LIWC categories and they suggest that this would, in turn, be a possible means to screen for dementia.

Asgari et al., (2017) examined the speech of people with mild cognitive impairment through a corpus-based and computational linguistic approach. They hypothesized that indicators of mild cognitive impairment may be present in people’s speech and such methods may be a way to predict further cognitive impairment. Asgari et al., (2017) used Linguistic Inquiry and Word Count to analyze their data. They explain that previous literature has “identified characteristic early disruption of normative patterns and processing of speech and language in patients with
Alzheimer’s dementia as well as prodromal dementia states such as Mild Cognitive Impairment” (Asgari et al., 2017).

Their data comes from a randomized control behavioral clinical trial to examine the effect of unstructured conversation on cognitive functioning among older adults who either have normal cognitive abilities or Mild Cognitive Impairment (MCI). Their sample consisted of 14 participants with mild cognitive impairment and 27 participants with no cognitive deficits. Asgari et al., (2017) explain that even though Alzheimer’s dementia “involves the memory domain, linguistic ability is also clearly affected.” Aspects such as secondary verbs per utterance, percentage of clauses, percentage of right and left-branching, propositions per utterance, conjunctions per utterance, mean duration of pauses and standard phonation time have all been noted as being significantly different between healthy, older adults and adults with MCI and Alzheimer’s dementia. Asgari et al.,’s (2017) methodology involved automating the identification with MCI using the analysis of language narrative samples. In addition to Linguistic Inquiry and Word Count (LIWC), they also applied a support vector machine classifier on extracted linguistic features to distinguish between participants with mild cognitive decline versus participants with no cognitive deficits.

Asgari et al., (2017) classified patients with mild cognitive impairment with 84% accuracy, which was higher than the purported 60% accuracy. Asgari et al., (2017) describe linguistic analyses of spoken language as a potential “powerful tool” in distinguishing people with MCI and people with “intact cognition.”

Corpus methodologies are growing in popularity in various research fields – including in areas and subfields that take computational approaches to assessing cognitive decline. In some cases, computational approaches are being applied to corpus data to even predict cognitive
decline by analyzing linguistic data. The aim of the current paper is to contribute to the identification of what linguistic elements vary amongst healthy individuals and individuals with mild or moderate dementia. The current study replicates methods that are similar to Shibata et al., (2016) and Asgari et al., (2017) as the current paper uses a corpus-based approach and also uses Linguistic Inquiry and Word Count to analyze the data from a clinical group and a normal, healthy group of participants.

**Data**

The motivation for this paper was to compare the language of people affected by Alzheimer’s dementia alongside a healthy, control group. Using pre-existing corpus data is a convenient means to compare linguistic variations amongst a clinical group and a healthy group. Our data originates from the Dementia TalkBank Pitt Corpus. The TalkBank Pitt Corpus is an open-source corpus consisting of cross-sectional data. The sample consists of interview data from 3 groups of participants: a healthy, control group, a group of patients with mild dementia, and a group of patients with moderate dementia. The study consisted of 104 participants in the healthy control group and 208 participants with mild and moderate dementia. Participants in the original study were interviewed using a picture task: “The Cookie Theft Picture.” Participants were asked to describe the picture task to a clinical interviewer. Their responses were recorded and transcribed. The original data was gathered as a part of a protocol which was administered by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine. The acquisition of the original Dementia TalkBank data was supported by NIH grants NIA AG03705 and AG05133 to the University of Pittsburgh. To analyze the dataset, we use Linguistic Inquiry and Word Count (LIWC) to analyze the interview transcripts. The independent variables in this project are the LIWC categories. Table 1 (featured in Appendix 1)
lists the categories adapted from Pennebaker’s (2015) LIWC categories and subcategories. The dependent variables are the 3 different groups: the healthy, control group, the mild AD group, and the moderate AD group.

**Methods**

*The Cookie Theft Picture*

![Image of The Cookie Theft Picture](image)

**Figure 1**

*The Cookie Theft Picture (Cummings, 2019)*

Picture descriptions are a well-established means of assessing expressive language skills and are used widely by clinicians in management of clients with language disorders (Cummings, 2019). Such tasks can be used supplementally alongside other formal language assessments. Picture description tasks can also be used during therapy to elicit both naming and sentence production in a continuous sequence of speech. The pictures used in these tasks are generally black and white line drawings and depict scenes that are relevant to cultural experiences. “The
Cookie Theft Picture” of the Boston Diagnostic Aphasia Examination Third Edition depicts a mother and two children in a domestic setting. Figure 1 depicts this image. Clients are instructed to describe each aspect occurring in the picture. Clients may be prompted by the clinician if they are omitting certain details or aspects of the image. There are not generally time constraints on the exercise, and it is usually recorded for the clinician to review. “The Cookie Theft Picture” aims to examine cognitive linguistic impairments across 7 different features:

1. **Salience of Information:** In the image, less salient information would be the plants in the background of the image or the clothing that the 3 characters are wearing. Clients that do not have neurological impairments are aware of the reduced salience of such minute details and will not mention such details towards the end of the task. Essentially, “The Cookie Theft Picture” reinforces the distinction between information of high versus low salience through placing most of the background details in the image beyond a physical barrier - such as the kitchen window.

2. **Semantic Categories:** The examinees must be able to use words for the people or “animate entities” in the image such as mother, girl, or boy, and must be able to use the appropriate words for inanimate entities, such as plate or counter. There are also abstract concepts that can be described in the image, such as the notion that mother is seemingly “daydreaming” by gazing out the window.

3. **Referential Cohesion:** In this case, an examinee must introduce people and objects during the description but then be able to reference those same people and objects later in the exam through use of pronouns. The most efficient way for this to be achieved is for the examinees to use indefinite noun phrases the first time
they mention a person or object and then use the respective personal pronouns (he, she, or it) to reference the entity that was denoted by the noun phrase or the anaphoric reference. For example, referential cohesion could be achieved through the following anaphoric reference: “A woman is in front of the sink. She is washing dishes.”

4. **Temporal Relations:** The examinee must be able to give a full description of the scene in the image. For example, if the examinee is able to state that the sink is overflowing in the image because the woman/mother left the tap running, they are able to give a more fleshed out account of the image.

5. **Mental State Language:** Theory of Mind is the cognitive ability to attribute mental states to one’s own mind as well as to other people. Mental states consist of knowledge and beliefs. For example, someone without a neurological disease would be able to extrapolate abstract concepts from the separate acts occurring in the image. Someone without a cognitive deficit would be able to infer that the sink is overflowing in the image because the mother is “daydreaming.” If a person is able to infer abstract concepts from the image, then they would possess Theory of Mind skills.

6. **Structural Language Skills and Motor Speech Production:** Deficits in semantics, phonology, and syntax can be detected via picture description tasks. For example, individuals with aphasia may have difficulties locating certain words and may subsequently pause more often while describing the image. Issues with syntax may also include issues producing certain grammatical constructions or may avoid them altogether. Individuals may use declarative sentences and
avoid passive constructions. Phonological errors may include confusing phonemes and using extra words to essentially “talk around” a word they cannot remember. “The Cookie Theft Picture” ultimately helps clinicians assess motor speech production skills and the intelligibility of an individual’s connected speech.

7. **General Cognition and Perception:** Since executive function can also be an issue for individuals with dementia, cognitive skills such as planning, and organization can be compromised. That said, if cognitive skills are compromised, individuals with dementia may not be able to perceive all elements of the scene in a coherent or logical manner (Cummings, 2019).

*Linguistic Inquiry and Word Count (LIWC)*

Linguistic Inquiry and Word Count (LIWC) analyzes text through a psychometric approach. Martha Francis and James Pennebaker created the software; a text analysis program that counts words in psychologically meaningful categories. Pennebaker writes that “the words people use in their daily lives can reveal important aspects of their social and psychological worlds” (2003). Furthermore, text analysis allows researchers to assess the features of what people say and the subtleties in their linguistic styles. Empirical results from various studies using Linguistic Inquiry and Word Count have demonstrated an ability to detect meaning in a wide array of experimental settings. These include attentional focus, emotionality, social relationships, thinking styles, and individual differences (Tausczik and Pennebaker 2010). In fact, the original goal of Linguistic Inquiry and Word Count was to calculate the percentages of negative and positive emotion words within a text. LIWC dictionaries are available in multiple languages. LIWC includes more than 2500 words, or word stems, categorized into groups of
words that are referred to as word “subcategories.” The subcategories fall into five broad “word categories:” 1) Linguistic dimensions, 2) psychological processes, 3) relativity, 4) personal concerns, and 5) spoken categories (Asgari, et al., 2017).

Certain categories in LIWC such as analytical thinking, clout, and mazes may appear somewhat abstract and must be explained for interpretation of later results. Analytical thinking is described as a “metric of logical, formal thinking” (Pennebaker et al., 2014). This variable is a factor-analytically derived dimension that is based on several categories of function words. Analytical thinking is meant to conceptualize the degree that people’s language reflects formal, logical, and hierarchical thinking patterns. In the case of LIWC, clout refers to either relative social status, confidence, or leadership that people may display in their language (Kacewicz et al., 2014). Additionally, we combined the LIWC subcategories of filler and nonfluencies to quantify mazes. Fillers may include utterances such as “you know” or “I mean,” while nonfluencies include utterances such as “uh, um, err,” etcetera.

Results and Discussion

Control group vs mild AD group

Figure 2 shows the differences between the control group and mild AD group across four subcategories: 1) analytical thinking, 2) use of the present tense, 3) nonfluencies, and 4) pronoun usage. We find that the control group uses more language that reflects analytical thinking than the groups with dementia. The control group also used the present tense in their speech more than the group with mild dementia. The group with mild dementia used more nonfluencies than the control group. Additionally, the group with mild dementia used more pronouns than the control group.
The higher use of nonfluencies on behalf of the mild AD group corresponds with our hypothesis that participants with AD would have issues with word retrieval would then use more pauses or filler words. The higher use of pronouns also corresponds with issues with word retrieval as well – with the implication being that as opposed to using the terms such as “brother, sister, or mother” participants used pronouns instead. The implication of the control group’s higher use of analytical language may reflect the overall ability to better articulate the picture task scenario than the two AD groups. The higher use of present tense on behalf of the control group may also indicate an overall ability to better articulate the scene in picture task as participants describe the image in real time to an interviewer.

**Figure 2**

*Control group versus Mild AD group*
**Control versus moderate AD group**

Figure 3 shows the comparison of the control group and moderate AD group across 4 subcategories: 1) words greater than six letters, 2) usage of articles, 3) usage of pronouns, and 4) usage of impersonal pronouns. We find that the control group uses more words that are greater than six letters. We also find that the control group uses more articles in their speech than the control. We also find that the moderate AD group uses more pronouns than the control group and that the moderate AD group uses more impersonal pronouns than the control group. The higher use of pronouns on behalf of the moderate AD group corresponds with word retrieval difficulties (Kempler and Goral, 2008). Pronouns may be more often used in place of nouns to compensate for the forgetting of a noun such as “mom or woman,” “brother or boy,” and “sister or girl.” This concept also applies to the moderate AD group’s higher use of impersonal pronouns. There is a higher usage of the pronouns “she” and “he” in place of familial terms. Higher usage of impersonal pronouns also corresponds to findings of Shibata et al.,’s 2016 study wherein the group with dementia used significantly more impersonal pronouns compared to the control group. The control group’s higher usage of articles – corresponds to their overall higher word count and higher words per sentence score than both the mild AD and moderate AD groups.

Additionally, the control group’s higher usage of words greater than six letters aligns with the general language trend of the control group having a higher word count and using more words per sentence than the two AD groups. Words greater than six letters tend to be indicative of language that is more complex (Pennebaker and Tausczik, 2010). The implication of higher usage of words greater than six letters may also relate to the control group’s overall stronger cognitive abilities and their ability to retrieve words.
Figure 3

Control group versus moderate AD group

Mild AD group versus moderate AD group

Figure 4 shows variations between 4 subcategories for mild AD and moderate AD: 1) clout, 2) relative, 3) pronouns, and 4) function words. We find that the mild dementia group uses language that indicates a higher usage of “clout” than the moderate dementia group. Clout as a LIWC category refers to language that is certain or confident or may reflect leadership. Higher usage of clout in the mild AD group may imply that the participants with mild AD seemed to speak in a way that is more confident or certain of their observations than the moderate AD group.

We also find that the group with mild dementia scored higher on use of language with references to relatives (i.e., family members). We find that the group with moderate dementia
uses more pronouns than the mild dementia group. We also find that the moderate group uses more function words than the mild group.

Use of the terms that correspond to family and relatives corresponds to the picture task itself but the higher use may indicate that the mild AD group is easier able to retrieve familiar terms (mother, sister, brother) Conversely, the higher use of pronoun usage in the moderate AD group may indicate that retrieval for relative related terms was more difficult for participants with moderate AD than for participants in the mild AD group.

Higher function word usage may also indicate a difficulty with specific word retrieval. This is because pronouns are included in the function word category, and we found that pronoun usage was higher in the moderate AD group than the mild AD group.

![Figure 4](image)

**Figure 4**

*Mild AD group versus moderate AD*
**General language trends**

The majority of our findings relate to presumed difficulties with word retrieval that is a key issue in cognitive decline. We find that the control group had the highest word count out of the sample, having a higher word count than the moderate AD group and followed by the moderate AD group. For words per sentence, we find that the control uses the most words per sentence. The mild AD group and moderate AD group both use fewer words per sentence than the control group. Regarding usage of function words, we find that the moderate AD group used the highest amount of function words. To calculate mazes, we combine the subcategories of *fillers* and *nonfluencies*. We find that participants with mild AD had speech with more linguistic mazes and with the control group having the fewest occurrence of mazes in their speech. We also take positive and negative emotion into account. We find that the control group uses the most amount of positive emotion in their language and that the moderate AD group uses the most amount of negative emotion in their language.

The control group has the highest word count as well as the highest amount of words used per sentence. This corresponds with the notion that people without abnormal cognitive decline have an easier time recalling particular words and overall having more cohesion in their language. The moderate AD group used the highest amount of function words. This corresponds with the group’s higher usage of pronouns, including personal pronouns, in the moderate AD group’s speech. Mazes is the combination of nonfluencies and fillers. Interestingly, the mild AD group’s language reflected a higher use of mazes. This is surprising given that retrieval of words and phrases becomes more difficult as Alzheimer’s dementia progresses.

Although we find trends in the usage of positive and negative emotion within the dataset, it is important to note that the emotionality reflected in this corpus specifically relates to the way
that participants describe the image, not necessarily personal one’s outlook being positive or negative. Emotions in the case of this generally relate to the descriptors of the image. For example, when describing the image one participant in the control group (female, age 69) says, “and the landscaping is bushes up close and happy evergreens and stuff.” Additionally, one instance of negative emotion is reflected by one participant in the mild AD group (female, age 70) as she says, “her water has spilled over terribly bad,” when describing the mother at the sink doing dishes. Figure 5 illustrates general language trends across 1) word count, 2) words per sentence, 3) function words, 4) mazes, 5) positive emotion, and 6) negative emotion.

![Figure 5](image)

**Figure 5**

*Language trends across all 3 participant groups*
Conclusions and future directions

A meaningful way to observe cognitive decline is by examining peoples’ language. Subsequently, corpus-based and computational linguistic methodologies are growing in usage in research areas that assess and treat cognitive decline. The current project used Linguistic Inquiry and Word Count (LIWC) to examine categorical differences in the language of people in a healthy control group and two clinical groups affected by mild and moderate Alzheimer’s dementia. Although Linguistic Inquiry and Word Count was an adequate starting point for linguistic analyses and can illustrate clear differences in the language of a clinical group versus a healthy group, it would be preferable to analyze the same sample data in a more robust manner. This would be done via a more in-depth methodology such as using machine learning algorithms to learn linguistic biomarkers, a methodology used by that of Orimaye et al., (2017).

As stated prior, LIWC is a good starting point for linguistic analyses, but it would be ideal to use a software that can analyze language further past the point of categorization and subcategorization. A limitation of this study is the necessity of human interpretation on behalf of the research which may cause subjectivity of interpretation of results (E.g., the use of positive versus negative emotion and how that is conveyed within the corpus itself).

In conclusion, the concept of using computational linguistic analyses to predict cognitive decline is growing in popularity. More research studies are needed alongside more robust analytical methods to possibly, and hopefully, predict the onset of cognitive impairment and Alzheimer’s dementia through linguistic analyses.
References


https://doi.org/10.31887/DCNS.2003.5.1/jloebachwetherell

### Table 1
*Categories and subcategories of LIWC dimensions, Adapted from Pennebaker et al., (2015)*

<table>
<thead>
<tr>
<th>Category</th>
<th>Abbrev</th>
<th>Example/Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word Count</strong></td>
<td>WC</td>
<td>Amount of words spoken</td>
</tr>
<tr>
<td><strong>Summary Language Variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytical Thinking</td>
<td>analytic</td>
<td>Metric of logical, formal thinking</td>
</tr>
<tr>
<td>Emotional Tone</td>
<td>tone</td>
<td>Degree of positive/negative tone</td>
</tr>
<tr>
<td>Clout</td>
<td>clout</td>
<td>Language of leadership/status</td>
</tr>
<tr>
<td>Words &gt; 6 letters</td>
<td>sixlitr</td>
<td></td>
</tr>
<tr>
<td>Dictionary words</td>
<td>dic</td>
<td>Percent words captured by LIWC</td>
</tr>
<tr>
<td><strong>Linguistic dimensions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total function words</td>
<td>funct</td>
<td>it, to, no very</td>
</tr>
<tr>
<td>Total pronouns</td>
<td>pronoun</td>
<td>I, them, itself</td>
</tr>
<tr>
<td>Personal pronouns</td>
<td>prpron</td>
<td>I, them, her</td>
</tr>
<tr>
<td>1st pers singular</td>
<td>i</td>
<td>I, me, mine</td>
</tr>
<tr>
<td>1st pers plural</td>
<td>we</td>
<td>We, us, our</td>
</tr>
<tr>
<td>Auxiliary verbs</td>
<td>auxverb</td>
<td>Am, will, have</td>
</tr>
<tr>
<td>Prepositions</td>
<td>prep</td>
<td>To, with, above</td>
</tr>
<tr>
<td>Common adverbs</td>
<td>adverb</td>
<td>Very, really</td>
</tr>
<tr>
<td>Negations</td>
<td>negate</td>
<td>No, not, never</td>
</tr>
<tr>
<td><strong>Other grammar</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common verbs</td>
<td>verb</td>
<td>Eat, come, carry</td>
</tr>
<tr>
<td><strong>Psychological Processes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affective processes</td>
<td>affect</td>
<td>Happy, cried</td>
</tr>
<tr>
<td>Positive emotion</td>
<td>posemo</td>
<td>Love, nice, sweet</td>
</tr>
<tr>
<td>Category</td>
<td>Subcategory</td>
<td>Example</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Negative emotion</td>
<td>negemo</td>
<td>Hurt, ugly nasty</td>
</tr>
<tr>
<td>Social processes</td>
<td>social</td>
<td>Mate, talk, they</td>
</tr>
<tr>
<td>Family</td>
<td>family</td>
<td>Daughter, dad, aunt</td>
</tr>
<tr>
<td>Male references</td>
<td>male</td>
<td>Boy, dad, his</td>
</tr>
<tr>
<td>Insight</td>
<td>insight</td>
<td>Think, know</td>
</tr>
<tr>
<td>Causation</td>
<td>cause</td>
<td>Effect, because</td>
</tr>
<tr>
<td>Differentiation</td>
<td>differ</td>
<td>Hasn’t, but, else</td>
</tr>
<tr>
<td>Perceptual processes</td>
<td>percept</td>
<td>Look, heard, feeling</td>
</tr>
<tr>
<td>Feel</td>
<td>feel</td>
<td>Feels, touch</td>
</tr>
<tr>
<td>Biological processes</td>
<td>bio</td>
<td>Eat, blood, pain</td>
</tr>
<tr>
<td>Body</td>
<td>body</td>
<td>Cheeks, hands, spit</td>
</tr>
<tr>
<td>Drives</td>
<td>drives</td>
<td>n/a</td>
</tr>
<tr>
<td>Achievement</td>
<td>achieve</td>
<td>Win, success, better</td>
</tr>
<tr>
<td>Power</td>
<td>power</td>
<td>Superior, bully</td>
</tr>
<tr>
<td>Reward</td>
<td>reward</td>
<td>Take, prize, benefit</td>
</tr>
<tr>
<td>Time orientation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past focus</td>
<td>focuspast</td>
<td>Age, did, talked, had, been</td>
</tr>
<tr>
<td>Future focus</td>
<td>focusfuture</td>
<td>May, will, soon</td>
</tr>
<tr>
<td>Relativity</td>
<td>relativ</td>
<td>Area, bend, exit</td>
</tr>
<tr>
<td>Motion</td>
<td>motion</td>
<td>Arrive, car, go</td>
</tr>
<tr>
<td>Personal concerns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leisure</td>
<td>leisure</td>
<td>Cook, chat, movie</td>
</tr>
<tr>
<td>Home</td>
<td>home</td>
<td>Kitchen, landlord</td>
</tr>
<tr>
<td>Informal language</td>
<td>informal</td>
<td></td>
</tr>
<tr>
<td>Netspeak</td>
<td>netspeak</td>
<td>Btw, lol, thx</td>
</tr>
<tr>
<td>Nonfluencies</td>
<td>nonflu</td>
<td>Er, hm, umm, oh, I-I</td>
</tr>
</tbody>
</table>