Automatic Tracking of Linguistic Changes for Monitoring Cognitive-Linguistic Health

by

Shuai Wang

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2016 by the
Graduate Supervisory Committee:

Visar Berisha, Chair
Amy LaCross
Hanghang Tong

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Many neurological disorders, especially those that result in dementia, impact speech and language production. A number of studies have shown that there exist subtle changes in linguistic complexity in these individuals that precede disease onset. However, these studies are conducted on controlled speech samples from a specific task. This thesis explores the possibility of using natural language processing in order to detect declining linguistic complexity from more natural discourse. We use existing data from public figures suspected (or at risk) of suffering from cognitive-linguistic decline, downloaded from the Internet, to detect changes in linguistic complexity. In particular, we focus on two case studies. The first case study analyzes President Ronald Reagan's transcribed spontaneous speech samples during his presidency. President Reagan was diagnosed with Alzheimer's disease in 1994, however my results showed declining linguistic complexity during the span of the 8 years he was in office. President George Herbert Walker Bush, who has no known diagnosis of Alzheimer's disease, shows no decline in the same measures. In the second case study, we analyze transcribed spontaneous speech samples from the news conferences of 10 current NFL players and 18 non-player personnel since 2007. The non-player personnel have never played professional football. Longitudinal analysis of linguistic complexity showed contrasting patterns in the two groups. The majority (6 of 10) of current players showed decline in at least one measure of linguistic complexity over time. In contrast, the majority (11 out of 18) of non-player personnel showed an increase in at least one linguistic complexity measure.

ACKNOWLEDGMENTS

This Master thesis summarizes findings of my recent work under the guidance of several professors. During my adventure to the realm of text analysis and language processing, I am very fortunate to get help from many people.

I would like to thank my thesis committee: Prof. Visar Berisha, Prof. Hanghang Tong, and Prof. Amy LaCross. They provide invaluable help with their expertise to help me finish this thesis.

I would also like to thank Prof. Visar Berisha, Prof. Julie Liss and Prof. Michael Dorman. I cannot make this thesis be possible without their recommendations.

Finally, I would like to thank my wife Danni Wang. She gives me a lot of support during this busy year.

Shuai Wang

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Cognitive-Linguistic Changes as Early Signs

Spontaneous communication is a cognitively demanding task that challenges memory function. Speech and language production requires simultaneous use of multiple cognitive resources, including memory, motor control, and attention (McClelland et al., 1986; Liberman and Mattingly 1985). When different modalities are involved, such as both acoustic and visual information, this challenge grows (Soto-Faraco et al., 2012).

Speech and language measures are an appealing metric for monitoring cognitive decline because they are sensitive to changes in cognitive state and because they are easy to collect. Several studies have documented that linguistic related impairments are early indicators of several of neuro-degenerative disorders, such as Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), and traumatic head injury (TBI) (Alzheimer's Association, 2014; Borgaro et al., 2003; Snowden et al., 1996). In addition, with the proliferation of mobile devices, speech and language data can be easily collected longitudinally (Stamford et al., 2015). Thus long-term changes in language production can provide useful information about the onset or development of any underlying neurogenic disorder.

There are many examples of this in the literature. Hier et al. (1985) showed that dementia subjects' speech samples contained less lexical complexity than the controlled normal subjects. Complexity was measured by counting the total number of words, the unique words, the number of prepositional phrases and other related metrics. As the patient progressed through advanced stages of dementia, the authors showed that they used more empty nouns and filler words (Hier et al., 1985). In another study focusing on AD, Bates et al. (1995) reported that the lexicon was gradually reduced through the development of AD. Bird et al. (2000) found that the decline in the proportion of suitable nouns and verbs that fit for the task was correlated with the severity of semantic dementia. Dodge et al. (2014) found that the proportion of words in free conversational speech can distinguish MCI group and AD group while other factors such as age and gender were controlled.

From the previous literature, it is clear that changes of linguistic complexity patterns often with neurogenic disease onset. Collecting longitudinal data in controlled experiments to confirm or extend these studies can be a time and resource consuming activity; However, there are a number of data sets online from individuals that we know eventually developed a neurogenic disorder that affects language. In this thesis, I will focus on exploratory analysis of speech and language measures extracted from these databases for two types of disorders: Alzheimer’s Disease, which occurs most frequently in the elderly population; and Traumatic Brain Injury resulting from concussion. The latter is often linked with injuries that occur during sport games or practices, especially among young athletes (Duma & Rowson, 2014; McKee et al., 2014). These two types of neuro-pathological disorders cover a wide range of age groups.

Background on Alzheimer’s Disease and Traumatic Brain Injury

AD is a progressive neurological disorder that is often linked to memory loss (Alzheimer’s Association, 2014). AD is the most common type of dementia, which accounts for about 60% to 80% of dementia cases in the United States (Alzheimer’s Association, 2014). Common linguistic symptoms include difficulty in naming tasks and difficulty in remembering recent conversations during early stages of the disease (Alzheimer's Association, 2014). These symptoms may evolve into difficulty in communication at a later stage (Alzheimer's Association, 2014; Smith et al., 1989). Alzheimer's Disease disproportionately affects the elderly population. About 5.2 million American have AD, of which 5 million are older than 65 years of age (Alzheimer's Association, 2014).

In addition to progressive neuro-pathological disorders associated with aging, language related impairments can also occur in populations with a history of TBIs (e.g., athletes in Boxing and Football). Fazio et al. (2007) found that scores of verbal memory tasks had a significant effect on three groups of subjects: control group (normal), concussed- symptomatic groups, and concussed-asymptomatic groups. The concussed symptomatic group had the lowest score (Fazio et al., 2007). Marini et al. (2011) compared cognitive, linguistic and semantic abilities between a group of 14 severe TBI patients and a group of healthy subjects through a storytelling tasks. Marini et al. (2011) found that TBI patients had adequate semantic complexities in their discourse while the language processing abilities such as cohesion were significantly worse than the control group. The MTBI group also showed deficits on other linguistic tasks such as naming, narrative production (King et al., 2006; Tucker and Hanon, 2009).

Athletes in helmeted sports, especially in the football, cope with a much higher risk of suffering TBIs (McKee et al., 2014), with concussions occurring during games and practice (Beckwith et al., 2013). In a 6-year study (1996-2001) following repeated concussions in NFL players, 887 concussions were reported in 650 players in practices and games (Pellman et al., 2004). Of the total number of participants, 160 players were associated with repeated head injuries and 51 players had three or more occurrences during the study (Pellman et al., 2004).

It is important to note that exposure to the risk of repeated head impact such as concussion is not limited to professional level football, but also college level, youth football and even elementary level football (Cobb et al., 2013; Nowinski 2006; Daniel et al., 2012). Additionally, those occurrences of head impacts in athletes that are below the clinical diagnosis level of concussion (called "sub-concussion") may also disrupt normal neurological functions (Bailes et al., 2013; Poole et al., 2015).

Motivation for Using Linguistic Markers

There are many reasons to prefer linguistic biomarkers. They may be indicators of early signs of dementia or other cognitive decline, speech and language data is sensitive to changes in cognitive health, and speech and language data is abundant and easy to collect. Below we describe the motivation in more detail.

Currently there is no cure for AD (Roberts & Petersen, 2014; Alzheimer's Association, 2014). However, numerous studies have shown that early detection of the disease can significantly improve outcomes in the future (Roberts & Peterson, 2014).

As with AD, accurate and early diagnosis of MTBI can improve outcomes in the future. Nordstrom et al. (2014) results indicate that early-life TBI may result in a much higher risk of obtaining serious dementia later in life. However, diagnosis of mild TBI can be very difficult because the majority of mild TBI happened without immediate symptoms such as loss of conscious (LOC) (Bailes et al., 2013). In sports, the early detection of MTBI also may help players at risk to prevent serious problems affecting their life post career (Guskiewicz et al., 2005). It is hypothesized that repeated, undiagnosed head trauma results in Chronic Traumatic Encephalopathy (CTE) (Mez et al., 2013; Omalu et al., 2010; Schwarz, 2010).

Speech and Language Measures Are Sensitive

A number of studies have shown that linguistic changes are able to capture signs of deterioration of cognitive health, often well before official diagnosis. Low performances on semantic and syntactic complexity analysis in written text at an early age (< 30 years old) were associated with confirmation of AD at later life (>70 years old) (Snowdon et al., 1996). Heitkamp et al. (2015) reported that linguistic analysis captured decline in linguistic complexity and linguistic variety of a man’s written diary in up to 7 years before he was diagnosed with semantic variant primary progressive aphasia. President Reagan, one of the more well-known AD patients, was not diagnosed with AD until 1994, three years after his presidential appointment ended. However, researchers suspected that he may have started showing symptoms earlier in life (Gottschalk et al., 1988). Le et al. (2011) analyzed texts of novels published by three prolific novelists and found early signs in measures of linguistic complexity and syntactic complexity in novels by one who has died with AD and one with suspected AD.

Speech and language data is abundant and easy to collect. People generate a great deal of speech and language data every day. Mehl et al. (2007) estimated that both women and men say about 16,000 words a day. This makes speech and text data an appealing passive tool to monitor cognitive health. Furthermore, this data is easy to collect compared to other expensive biomarkers such as the proportion of protein fragments currently used in the diagnosis of AD (Dodge et al., 2014). For example, written messages in emails or text messages can be collected in background applications.

This means that we can perhaps use these data sources to monitor statistically significant longitudinal changes in language that are consistent with expected trends for different disorders.

Proposed Study to Monitor Cognitive-linguistic Decline Using Publicly Available Data

The majority of previous studies that have analyzed changes in speech and language in dementia have focused on carefully collected data. The principal aim of this thesis is to see whether these measures of linguistic complexity are also sensitive to changes in unscripted, uncontrolled conversations. To that end, I focus on public figures that are suspected of dementia or at risk for CTE for which there exists a great deal of publicly-available language data online, because public figures often have a lot of records of their transcripts already online. So transcripts from public figures who are confirmed with AD or with MTBI can be used for retro perspective analysis.

Why publicly available data? There is a great deal of publicly available data generated every second online (Huberman and Adamic, 1999), with a large portion of these unstructured data sources containing either speech or language samples. Examples of these data include archives of interviews, talk shows, online videos, etc. These data sources are also often tagged with additional meta information that can be useful for health applications. For example, researchers have successfully merged social media files such as Facebook or Twitter with corresponding Electric Medical Record (EMR) data to find underlying information that is helpful for general disease diagnosis (Padrez et al., 2015).

We can mine these existing large speech and language data sources to detect changes in language complexity. Changes in language patterns are slow (Le et al., 2011; Burke & Shafto., 2008). As a consequence of slow changes in language patterns, collecting language data longitudinally in controlled behavior experiments is a difficult and resource-consuming task. As a result, these experiments often only include a few subjects. For example, some previous longitudinal studies only involved either a picture description task or a narrative task done annually in only a few subjects (Bird et al., 2000; Kemper et al., 1989; Kemper et al., 2001; King et al., 2006). Finally, psycholinguistic experiments require a fine control in the set up of experimental parameters in order to get over the issue of the small sample size. The experiment design further narrows down the possible linguistic patterns to assess from data. Some linguistic complexity measures (e.g., type-to-token ratios) are often based on proportions or probabilities, which requires sufficiently large data to be consistent.

In summary, the massive speech and language data available online provide us with a unique opportunity to examine linguistic trends in individuals suspected of.an underlying disorder. Results from massive online data can go far beyond traditional psychology behavior experiments. This massive archive of internet data even makes us possible to track data years back.

The present study describes two collected data sets from publicly available transcript data. The study investigated possible language markers for early signs of cognitive impairments such as AD and TBI. Results showed the potential of using massive data mining technique and publicly available data sources, to capture subtle changes in cognitive-linguistic measures.

Because I relied on retrospective analysis of publicly available data, one of the major issues to deal with is noise in the data. In this thesis, I will discuss how I clean and smooth the raw data directly from the Internet without hindering the power of finding long term linguistic trend. Two case studies used transcribed answers from public figures as data source to analyze longitudinal linguistic patterns.

The first study is a case study on the transcripts of news conferences from two former Presidents: President Ronald Reagan and President George HW Bush. We analyzed transcripts of their spontaneous answers to questions from the press. Results showed that Reagan had a steady decline in linguistic complexity while Bush didn't show any trends under the same analysis. Reagan was diagnosed with AD in 1994 while Bush has never been diagnosed with AD.

The second study expanded and generalized the ideas of the first study on a much larger scale. We collect data from tens of thousands of webpages of news conferences from 10 NFL players and 18 non-player personnel from the year 2007 to the year 2015 and applied a similar analysis as the first study to this NFL data set. Results showed contrasting patterns of linguistic complexity between the group of non-player personnel and the group of players. A considerable percentage of players had decline in measures of linguistic complexity while most of the non-player personnel showed increase in the same set of linguistic complexity measures.

Results from two case studies encourage the general idea to take advantage of current mobile technology and social media to automatically monitor the cognitive-linguistic health status for groups of interest, such as athletes who are at risk of TBI, or those that are predisposed to dementia.

Contributions of This Thesis

There are two principal contributions to this thesis: (1) publications that describe the results of my analysis; (2) Python scripts for analyzing linguistic complexity directly from transcript data.

Papers Published and Presentation:

Visar Berisha, **Shuai Wang**, Amy LaCross, Julie Liss. (2015). *Tracking Discourse Complexity Preceding Alzheimer's Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush*, Journal of Alzheimer's Disease (impact factor: 4.15), *45,* 959-963.

Submission in preparation: Visar Berisha, **Shuai Wang**, Amy LaCross, Julie Liss. (2016). *Changes in Linguistic Complexity in Professional Athletes At Risk for Chronic Traumatic Encephalopathy.*

Script Framework Developed:

We have started writing a toolbox for measuring linguistic complexity form publicly available data. Currently we have a collection of scripts that can be used to: (1) download data directly from web address. (2) Converting webpages to text files with cleaning. (3) Communication between linguistic analyzer and local storage (4) basic linguistic feature extraction such as classes of words. The future plan is to organize those scripts into a Python package.

CHAPTER 2

REVIEW ON LINGUISTIC FEATURES USED IN MEASURING COGNITIVE-LINGUSTIC DEFICITS

Introduction

Oral and written communication in the English language is organized hierarchically. Hundreds of morphemes are used to compose thousands of words (e.g. *have),* along with prefixes such as *en* and *un*, and suffixes such as *ing* and *ed*. Word sequences compose phrases that convey meaningful information and those sequences follow grammatical criteria to compose more complex sentences. Structuring these sentences is a cognitively-taxing task that requires a complex combination from multiple regions in the brain (i.e., motor theory of speech perception from Liberman and Mattingley, 1985). A number of studies have shown that the complexity of spoken and written text decreases with neurological disease progression (Snowden et al., 1996; Le et al., 2011; Heitkamp et al., 2015). In this thesis, we posit that this is especially true in spontaneous discourse where speakers have to construct answers in real-time.

Studies of AD, MTBI, and normal aging control groups have used numerous linguistic complexity measures (Bird et al., 2000; Cheung & Kemper, 1992; Hier et al., 1985; Kemper et al., 1989; Kemper et al., 2001; Le et al., 2011; Snowden et al., 1996). In this chapter we review a series of lexical/semantic features that can be measured directly from language samples (e.g. transcribed speeches) to estimate the complexity of the language. It is a subset of these features that we track longitudinally in both case studies in Chapters 3 and 4. These groups of features measure two complementary aspects of language: 1) lexical complexity measures based on word-level information, and 2) syntactic (grammatical) complexity measures based on sentence-level information. We describe both measures of complexity.

Lexical Complexity Measures

Lexical complexity describes the level of "meaningfulness" in a transcript sample. In lexical complexity measures, words in the original text need no further segmentation or decomposition into roots. Instead, lexical complexity measures use word-level information directly. Almost all lexical complexity measures involve categorizing words from a collection of words. Some lexical complexity measures such as type-to-token ratio have a nonlinear relationship with text length (Le et al., 2011), so normal semantic measures are comparable when a fixed size is set for the total amount of words (or tokens). Below we describe the list of semantic features.

Type-to-token ratio (TTR). The type-to-token ratio, a proxy for vocabulary size, is one of the most commonly used measures. Here, *type* refers to words after lemmatization, in which suffixes or prefixes are removed and thus change the primitive properties rather than the meanings of words. For example, after lemmatization, *books* become *book*. The TTR measures can be further separated into language units such as noun TTR, verb TTR, adjective TTR, and adverb TTR (Le et al, 2011).

Word–class deficit. One way to categorize words is based on their basic grammatical functions, such as nouns, verbs, adjective, adverbs, and prepositions. Besides statistically based TTR measures, word–class distributions can also reflect changes in lexical complexity (Le et al., 2011). An onset-dementia study of a group of dementia sufferers showed significantly declining portions of noun-class tokens and significantly increasing portions of verb-class tokens (Bird et al., 2000; Le et al., 2011).

Filler words. Speakers often use filler words to connect thoughts or for momentary pauses. Increased use of filler words was shown during the development of severe dementia (Hier et al., 1985). We use a statistical count of lexical filler words: *well*, *so*, *basically*, *literally*, and *actually*. Other studies also used non-lexical filler utterances such as *uh*, *ah*, and *umm* (e.g., Kemper et al., 2011).

Lexical Density. In Lu (2012) 's lexical complexity analyzer, lexical density is defined as number of open set tokens such as nouns, verbs, adjectives and adverbs out of total number of tokens. The ratio of open set tokens is also a proxy for the vocabulary size.

Idea Density. This measure is defined as the number of ideas expressed through verbs, adjectives, adverbs, or prepositional phrases in every 10 words and is related to factors such as educational levels and vocabulary size (Snowdon et al., 1996). The definition of idea density is similar to the definition of lexical density (see details in case study II methods session) in lexical complexity analyses (Lu, 2012).

Lexical repetition. Global repetition refers to frequent use of certain *n*-grams (*n*-word sequence). More frequent use of the same word combinations indicates reduced vocabulary. Local repetition refers to the portion of all open-set nouns, verbs, adjectives, and adverbs after lemmatization. Global and local repetition examples can be estimated directly from written samples (Le et al., 2011).

Non-specific nouns. Concrete nouns identify the subject being discussed, but non-specific nouns lack specific subject reference. Increased use of non-specific nouns indicates reduced vocabulary (Le et al., 2011; Nichola et al., 1985). Here, we identify non-specific "thing" nouns such as *nothing*, *anything*, *something*, and *things*.

Low-image verbs. Low-image verbs are similar to non-specific nouns in that the verbs lack specific actions or precision. We use 14 verbs from Bird et al. (2000): *be*, *com*e, *do*, *get*, *give*, *go*, *have*, *know*, *look*, *make*, *see*, *tell*, and *think*. These modal verbs often combine with other verbs to make meaningful sentences such as “I have received your order.” Decreased use of specific verbs indicates a low complexity in semantic information; increased use of low-image verbs also indicates reduced vocabulary.

Syntactic Complexity Measures

MLU. MLU, *length per utterance*, is a measure of sentence length based on the number of words (Cheung & Kemper, 1992).

MCU. MCU, *mean number of clauses per utterance*, is measured by counting the number of main and sub-clauses (Cheung & Kemper, 1992; Le et al., 2011). i.e., *what you have is good to believe* has a main clause (*something is good to believe*) and a sub-clause (*what you have*). Sentences with more than one clause are more complicated compared to sentences that only have one clause (*Today is a good day*).

Development Sentence Scoring (DSS). The DSS was developed to assess childhood development of grammatical structure (Lee, 1974). The DSS score is based on the use of words indicating differences such as indefinite nouns, personal nouns, negative words, and secondary verbs (Cheung & Kemper, 1992).

Developmental Level (DLevel). Developmental level refers to *grammatical complexity* (Snowdon et al., 1996). The original DLevel is a manual classification scale ranking sentences from simplest to most complex: 0 for sentences with only one clause; 7 for sentences with multiple sub-clauses (Rosenberg and Abbeduto, 1987). The levels were later extended from seven to eight levels (Cheung and Kemper, 1992).

Directional Complexity (DComplexity). A scale similar to the DLevel was developed for measuring the syntactic complexity of sentences based on a relative readability complexity (Botel and Granowsky, 1972). The scale is from 0 to 3; for example, zero-count simple-structured sentences have only 2 or 3 lexical items (e.g., *He works here*). Three-count complex-structured sentences may have embedded clauses serving as subjects (e.g., *What he does is good*).

Passive Voice. The passive voice measure is based on the number of sentences containing be-passive, get-passive or by-passive constructions in the total number of sentences, but its utility as a complexity measure is unclear (Le et al., 2011).

Index of Productive Syntax (IPSyn). The IPSyn was developed to assess grammatical complexity in children less than 48-months-old (Scarborough, 1990). The measure is not based on syntactic complexity of individual sentences. Instead, it scores matched occurrences of 56 grammatical structures over a set of sentences (e.g., 100 utterances). Scores are further divided into subcategories such as noun phrases, verb phrases, and questions (Scarborough 1990).

Yngve Score. The Yngve score was developed for syntactic complexity scoring, based on the concept of limited working memory for processing sentences (Yngve, 1960). Speaking of a sentence requires information to process from left to right (Burke & Sharfot, 2008). For example, a person says *what he did today is not acceptable.* The clause *what he did* will come first. This clause has to stay in the working memory until the whole sentence finishes. A sentence like *it is not acceptable for what he did today* is easier because here the main left clause is just *it.* In that case, left-branching clauses require more cognitive resources because they must be held in memory until the sentence ends (Burke & Shafto, 2008). Sentences with longer left-branching clauses will have a higher Yngve score. Cheung & Kemper (1992) provided how to calculate Yngve Score in an example.

Frazier Score. Similar to the Yngve score, the Frazier score calculates the syntactic complexity according to sentence structure (Frazier, 1985). Frazier scoring assigns 1 to nonterminal nodes and 1.5 to terminal nodes in a parsing tree (Cheung & Kemper, 1992). The sum of scores along the path from surface to root give the score of each node. A higher word score indicates greater depth. The average Frazier score is the mean of scores of all words in a sentence (Cheung & Kemper, 1992).

Tools for Measuring Linguistic Complexity Directly from Text

We are currently developing a toolbox to measure complexity directly from text. Below we describe the existing tools and methods used to implement the syntactic and lexical complexity measures.

NLTK. The Natural Language Processing Toolkit (NLTK), written in Python, is one of the more popular natural language processing packages. The NLTK provides a large collection of both functions and databases related to natural language processing tasks including language modeling, stemming, part-of-speech (POS) tagging, and sentence parsing. These functions help decompose sentences into meaningful units for additional analysis. The NLTK also provides collections of databases and text corpora useful for common NLP tasks such as the CMU pronunciation dictionary, the Brown Corpus, and the Penn Tree Bank Corpus. Useful for implementing the measures previously described, the NLTK also includes the following functions:

Tokenization. Tokenization, a fundamental step for most of the complexity metrics, lists tokens in a sentence.

Stemming. Stemming removes prefixes or suffixes from words and converts them to their root words (e.g. using *book* for *books)*. The NLTK package provides the Lancaster stemmer for direct stemming tasks.

Parsing and POS-tagging. The NLTK has several parsers available depending on the underlying grammar structure used (Bird et al., 2009). Common parsers will return a tree-like string (embedded in parentheses) to indicate the surface structure of a sentence with tags such as "VP" (verb phrase).

Lexical Complexity Analyzer (LCA). In addition to using the NLTK for direct lexical complexity analysis, we can also use comprehensive linguistic complexity analyzer to get a matrix of standard linguistic complexity measures. LCA, developed by Xiaofei Lu and Haiyang Ai at Pennsylvania State University, (Lu, 2012) has a primary goal of assessing language development in the oral narratives of learners of English as a second language. However, LCA linguistic measures have many standard metrics identical to those we use in studies of written examples from dementia patients. The measures include TTR and lexical variations. The LCA is based on Stanford POS tagger for word level lexical complexity analysis and the Stanford Parser for sentence level syntactic complexity analysis. Here, we focus on word-level lexical complexity analysis. In addition to an offline software package, the Lexical Complexity Analyzer (LCA) also provides an online interface (http://www.personal.psu.edu/xxl13/downloads/lca.html) for extracting linguistic features through uploaded text files.

In the word-level lexical complexity analysis, online LCA analyzer will return a matrix of 25 linguistic features including statistical analysis such as word token count, TTR, lexical density, lexical variation, and lexical sophistication for each text file uploaded to the server. The full details of linguistic features and their definitions can be found in Lu (2012) for lexical complexity analysis and Lu (2010) for syntactic complexity analysis on sentence-level information.

Rationale of Using Lexical Complexity Measures

These lexical complexity measures and syntactic complexity measures have been used more or less in studies on patients with cognitive disorders who show language deficits (Hier et al., 1985; Kemper et al., 2001). These existing measures provide a candidate set of language measures that are useful for the present study. However, the present study only used lexical complexity measures.

Why do we only use lexical complexity measures? Here, I use only word-level semantic complexity measures for two reasons. First, it is known that signs of onset and development of dementia such as AD include the decline of semantic complexity such as type-to-token ratio and increased use of empty nouns, verbs, and fillers (Hier et al., 1985). However, normal aging groups often show similar declines in syntactic complexity because working memory shrinks (Burke & Shafto, 2008; Kemper et al., 2001). Consequently, longitudinal studies of syntactic complexity are difficult to interpret because normal aging and cognitive impairment have similar impacts on syntactic complexity measures.

Second, my data are transcribed spontaneous speech samples from sessions such as Q&A conversations. The transcripts contain many sentence fragments rather than full sentences. (See Table 2 in Case Study II for sample Q&A sessions.) The fragments will not work with standard parsers trained in formally written sentences.

We purposely chose to use web transcripts rather than prepared statements as the data source. Instantaneous responses to questions require more cognitive resources and are more likely to challenge cognitive–linguistic abilities.

CHAPTER 3

CASE STUDY: TRACKING DISCOURSE COMPLEXITY OF FORMER PRESIDENT RONALD REAGAN AND GEORGE H.W. BUSH

Introduction

Former President Reagan was diagnosed with AD in 1994. However, there was speculation that he had started showing signs of cognitive decline well before then (Gottschalk et al., 1988). Publicly-available presidential archives allow us to review changes in his linguistic abilities during the time he was in office. In this first case study, I reviewed his answers to questions in presidential press conferences from 1981 to 1988 - the time he was in the Oval Office.

I measured a subset of the many linguistic complexity measures we described in the previous chapter and found that both the unique word count, the number of filler words plus non-specific words were consistent with declining linguistic complexity. As a control, we compare and contrast Reagan's linguistic patterns with former President George H.W. Bush's news conferences from the same source using the same metrics. President Bush took office in 1989 (64 years old) and he has no known diagnosis of dementia. President Bush revealed no decline in linguistic complexity in his press conferences during the four years he was in office. In this chapter, we review the specifics of this study. Below I describe the methods, results, and discuss the findings.

Methods

General methodology for getting and cleaning the data. Figure 1 describes the general methodology for getting and cleaning the data in the case study. Following paragraphs will focus on explaining these steps in detail.

Materials. The American Presidency Project (http://www.presidency.ucsb.edu) has transcriptions of important documents from presidential history, including state of the union addresses, news conferences, and important speeches. For measuring linguistic complexity, we focus only on the news conferences of the two presidents, since they are in the form of a question-answer (Q&A) session and require a real-time response to questions. The spontaneous nature of the interaction makes this situation cognitively taxing. The transcript of news conferences usually starts with a prepared statement from the president. Following the statement, there is a Q&A session between journalists and Presidents. The Q&A sessions include questions from audience (typically journalists) followed by spontaneous answers from the President. The present study only focuses on spontaneous answers from the President because those answers require more cognitive resources than simply reading the initial prepared statements.

Data Processing. In the present study, answers from each President's news conference transcripts were downloaded and saved as a text file. The text was further cleaned by removing annotations, numbers. After the cleaning process, the first 1,400 words of each text were analyzed. The lower threshold of 1,400 words is determined from the shortest news conference from President Ronald Reagan. It's important to maintain a consistent word size for each sample because lexical measures are correlated with the length of the candidate text (Le et al., 2011). We stem the words via the Lancaster stemmer on NLTK. All 46 news conferences (from 1981 to 1988) from Reagan were used. 101 of 137 news conferences from Bush (from 1989 to 1993) were used because not all news conferences from Bush data provide at least 1, 400 words.

Text Statistics. We use the NLTK to write scripts to analyze the transcripts in order to obtain counts of unique words, non-specific nouns, filler words and low-image verbs. The non-specific nouns are words such as “thing”, “things”, and “something”. These nouns are often used for substituting a specific subject in the speech. Filler words are words such as “well”, “basically”, “so”, “actually”, and “literally”. The literature has shown that the increase of use of non-specific nouns and filler words are associated with the onset of dementia (Kemper et al 1989; Hier et al., 1985). The low-image verbs are 14 verbs defined in Bird et al (2000) to monitor the progressive degradation of language ability in semantic dementia.

Measuring longitudinal change. We use the correlation between the transcript index and the linguistic complexity measures to track changes in these parameters over the course of the 8-year period. A negative correlation coefficient for unique words implies a decline in complexity (e.g. reduced vocabulary); a positive correlation for the fillers and non-specific nouns implies an increase in the redundancy of language and a decrease in specificity.

Results

At the start of their presidential terms, President Reagan was 69 years old and President Bush was 64 years old. Table 1 contains a list of descriptive statistics for the analyzed press conferences. The comparison on average number of news conferences that two Presidents took showed a significant difference (unpaired $t = 31.3$, $p < 0.0001$, results derived from Berisha et al., 2015 with permission). President Bush used few unique words that President Reagan (unpaired $t = 2.6$, $p = 0.010$, results derived from Berisha et al., 2015 with permission). Bush used more low-image verbs than Reagan (unpaired $t = 7.8$, $p < 0.0001$, results derived from Berisha et al., 2015 with permission); Reagan used more filler words and non-specific nouns that Bush (unpaired $t = 3.9$, $p = 0.0001$, results derived from Berisha et al., 2015 with permission). These statistics showed a difference in the speaking style between Reagan and Bush. My interest in this work is to track longitudinal language changes in Reagan and Bush's data.

The correlation analysis between the transcript index and the linguistic complexity measures shows that count of unique words had a significant decreasing trend ($r = -0.446$, $p = 0.002$) with transcript indexes. The analysis also shows that count of non-specific nouns plus filler words per 1,400 words of each news conference showed a significant increasing trend with transcript indexes ($r = 0.358$, $p = 0.017$). For other linguistic measures no significant results were found (see Table 2, derived from Berisha et al., 2015 with permission).

Neither of the two metrics that reveal a significant trend for President Reagan show a significant trend for former President Bush's samples (See figure 1, derived from Berisha et al., 2015). President Bush was 5-year younger than Reagan, however he was the closest person we find to match Reagan's age. In order to match both their ages and number of transcripts, additional analysis on his last 46 transcripts over the last two years (when he was 66 years old) of his incumbency didn't showed significant trend in count of unique words ($r = 0.018$, $p = 0.369$) or count of non specific nouns +fillers ($r = 0.046$, $p = 0.150$) (these results were derived from Berisha et al., 2015 with permission). The result indicated that Reagan may show signs of cognitive-linguistic impairment, which was consistent with previous analysis of by Gottschalk et al. (1988) on president Reagan's debate data.

CHAPTER 4

CASE STUDY: TRACKING LINGUISTIC COMPLEXITY THROUGH NEWS CONFERENCES OF NFL PLAYERS AND NON-NFL EXECUTIVES

Introduction

On-field injury records and off-field assessments show that football players face a significant risk of concussion during their professional careers (Pellman et al., 2004; Benson et al., 2013). The impact of repeated head impacts may even exist long after their retirement (Omalu et al., 2005; Omalu et al., 2010). Guskiewicz et al. (2005) reported onset of AD in the group of retired American football players was earlier compared to American male population. In this case study, we analyze changes in linguistic complexity over time in professional football players from the national football league (NFL). We hypothesize that chronic head trauma and diagnosed concussions lead to overall decreased linguistic complexity. Similar to the first case study involving the former presidents, I analyzed transcribed spontaneous speech samples from 18 non-player personnel and 10 NFL players from 11 teams, spanning the years 2007 to 2015. Longitudinal data was collected using a web crawler followed by a pipeline of natural language processing. The TTR and lexical densities (LDs) were extracted from blocks of data for each person included in the study. Linear regression analysis was performed between time stamps of those text blocks and the corresponding linguistic measures.

By comparing linguistic changes between non-player personnel who have never played professional football and NFL players, this study found a decline in linguistic complexity for NFL players, but no such decline for non-player personnel who have never played in the NFL (Omalu et al., 2010). Below we describe the methodology of the study and describe the results.

Methods

Sources of data. Original articles were crawled from NFL team websites. Each NFL team maintains a website that posts news and updates periodically. Certain teams include "Transcripts" or "News Conferences" categories in the news website. A subset of these news has "Question & Answer" (Q&A) style transcripts that include interviews with the players. Using the Python Packages Selenium, Newspaper, NLTK, I downloaded the corresponding Hypertext Markup Language (HTML) webpages and processed them to extract the text and the date of the interview.

Although NFL teams have similar goals for the website layout, not all NFL teams provide Q&A style articles to download. I was able to identify 11 teams that provide such articles; Table 3 summarizes the names of those teams and provides links to sample articles.

After the general crawling process, articles were downloaded together with their original title and a corresponding date stamp at the time when it was published. The files were saved in a format that titles included the player's name and the date of the interview. Later time stamp information was extracted from the title. For example, "QB_Tom_Brady_News_Conference_Dec_1st_2012" is an interview with New England Patriots QB Tom Brady from Dec 1st 2012. Three volunteers manually inspected titles, names and content of these articles to validate the results of the crawling process.

Data processing. Original text files were made up of lines of text either answers from the interviewee or questions from the press. Occasionally, there were opening statements from the interviewee before the Q&A session begins – these were discarded. After processing of the data, only answers after questions were included in the processed text data for analysis. Additionally, words not from the interviewee, such as annotations within quotes (e.g. laughter), were removed. Answers from the same person were connected into a single text ordered by the dates of the interview.

Data cleaning. To further smooth the data, I applied a sliding window to the connected stream of text for each person included in the analysis. The sliding window picks a set of sentences that have just over 1000 words. The window will stop at where there are 1000 words. There is also a 30% overlap between two adjacent windows. This means that 30% of words will be included in the next sliding window stream. The 30% overlap smoothies the data. Each window of texts is exported to an independent text file with time stamps of that sliding window. Together, a stream of text files with equivalent size and monotonically increasing date stamps is created for each person. The date stamp of each block is determined by the published date of the first part of the text.

Inclusion criteria for data after cleaning. In order to make later statistical analysis valid, I set two inclusion criteria. The first one is that the candidate's data should be long enough for the longitudinal analysis. The minimum number of text blocks included in the study is 45 and each block contains 1000 words (size of the window). The second criterion is that personnel in the non-player group should have never played professional football in their life. This criterion ensures that a history of playing profession football will not be a confounding variable in later analysis. To date, there are 28 individuals from the crawled websites that satisfy these criteria. These include 10 players and 18 non-player personnel who have never played in the NFL. Demographic information such as name, interview date range, age and education can be found in Table 4 and Table 5. Together they represent 11 out of the 32 NFL teams.

Programming and analysis platforms. The data was processed on Python. Web crawling and transformation from HTML pages to text files was implemented using the packages Selenium, Newspaper, and NLTK. The processed text files are analyzed using the Penn State Linguistic Complexity Analyzer (Lu, 2012). This analyzer was developed based on the Stanford parser and POS-tagger on the Python platform. Current lexical complexity (word level) analysis is designed to reflect lexical variations and language sophistication, especially in the development of language complexity over a bank of linguistic measures (Lu, 2012).

Linguistic complexity measures. The previous study by showed that TTR and ratio of meaningless words reflected chronic changes in cognitive health by analyzing public speeches from former Presidents. Similarly, two metrics from the Penn State Linguistic Complexity Analyzer were retrieved for the analysis. The first metric is type-to-token ratio (TTR), which is the number of unique words they produce (and is an estimate of vocabulary size). For example, for the sentence "We lost because we scored less", unique types are *we, lose (root for lost), because, score, less.* There are total of 5 types*.* Tokens are *we, lost, because, we, scored, less*. There are total of 6 tokens. So the TTR=5/6=0.8. The second metric we use is the Lexical Density (LD), which is the ratio of content words (open set words such as nouns, full-verbs, adjectives, adverbs that have adjective base; opposed to grammatical words) to total number of words. For example, for the sentence "We lost because we scored less", the open set tokens are *lost, score, less*. There are 3 open set tokens. The total count of tokens is 6, same as the TTR example. So the LD=3/6=0.5. Both are positive metrics with range from 0 to 1. Higher values indicate higher lexical sophistication and higher lexical variation (Lu, 2012).

Results

Linear regression analysis on time stamps and linguistic complexity.

Statistical results are generated using the Analysis Toolbox in Prism 6. For each person's data, a linear regression analysis between the time stamps of text blocks and corresponding lexical density value and type-to-token ratio values are fitted with 95% confidence interval. The time stamps are calculated as number of days of since 2005/1/1. The *p*-value of the slope is also reported table 6 for linear regression fitting of LD and table 7 for linear regression fitting of TTR, respectively. The *r* values associated with the two variables in each regression (and the corresponding p-values) are also reported in Table 6 and Table 7. The Δ LD and Δ TTR for those showing either a significant increasing trend or a significant decreasing trend, are also reported in Table 6 and Table 7, respectively. These values are based on estimated values from linear fit regression equations. It is calculated as the difference between fitted intercept of ending block and intercept of starting block for each person's fitted regression slope and intercept value. To better view the analysis results, data which showed significant increasing trends are marked in green, while those showed significant decreasing trend are marked in red.

Starting points for TTR and LD between two groups. It is hypothesized that non-player executives and players perform equivalently on language complexity at the starting of their own longitudinal study, because they are all mature adults with equivalent education levels. In that case, I also reported the estimated starting point for both lexical density measures and type-to-token measures, together with the observed interview transcripts start date and end date in tables 4 and table 5. Similarly, these starting points for lexical density and type-to-token ratio values are estimated based on the intercept of the fitted equation from linear regression analysis and value of the starting block date. These results indicate that players and non-player start at the same level of lexical density measures and type-to-token ratios measures. Unpaired t-tests showed no significant difference in the mean of either lexical density ($p = 0.394$) measures or type-to-token measures ($p = 0.058$) between the group of players and the group of non-players (Figure 3).

Age comparison between two groups. An unpaired *t*-test shows that there is a significant difference between the age of non-player executives (mean=53.91 yrs old) and players (mean =31.30 yrs old) (see figure 4).

To summarize the results, 11 out of 18 from the non-player personnel group showed a significant increasing trend in at least one of two linguistic measures. 5 out of 18 from the group showed a significant trend in neither one out of two linguistic measures. Four out of 18 from the group showed a significant decreasing trend in at least one out of two linguistic measures.

Only 1 out of 10 players show a significant increasing trend in one out of two linguistic measures. Seven out 10 players showed a significant decreasing trend in at least one out of two linguistic measures. 3 out of 10 players showed no significant trend in either measure. As an illustrating example, for Tom Brady, both the LD and TTR decreases as a function of number of games played. This is shown in figure 5.

CHAPTER 5

CONCLUSION AND DISCUSSION

Compare Case study I with Other Studies

The decline of linguistic measures in the case study of Reagan's data is consistent with previous studies. These studies show that reduced lexical complexity was correlated with the onset and development of dementia. Smith et al. (1989) reported that Dementia of Alzheimer Type (DAT) patients produced less concise information on the same picture description tasks than control groups. DAT patients produced more "idiosyncratic utterances" than control groups. Idiosyncratic utterances refer to utterances that are grammatically correct but not semantically related to the task. As dementia becomes more severe, AD patients produced more empty nouns and filler words (Hier et al., 1985). The current study demonstrates that decline in linguistic complexity occurred when Reagan was still in the office. This finding was consistent with that of Gottschalk et al. (1988). Gottschalk et al. (1988) suspected that Reagan had deficits in verbal communication in public debates back to year 1984, during his second term as the President. In contrast to their study, which relied mostly on subjective evaluation using a standard test, my approach was based on objective measures of linguistic ability.

The analysis of Bush's data in case study I supports that lexical complexity can be preserved in normal aging. Kemper et al. (2001) showed that normal aging group could still maintain lexical and syntactic complexity until their mid 70s. Bush was 64 years old at the start of the analysis. No significant trend was found in lexical complexity analysis in Bush's data.

The Bird et al. (2000) study found that patients increased the use of low image verbs during the development of semantic dementia. In my study, we did not find any significant trends in the similar analysis from Reagan and Bush. This may be due to the small sample size (e.g., 46 for Reagan and 101 for Bush) in this study.

Comparison of Case Study II with Other Studies

Because case study II only includes 10 NFL players, there is not enough evidence to draw the connection between concussion and declines in linguistic complexity. However, two of the players (Andy Dalton, Robert Griffin III) had official NFL injury records in either "concussion" or "head injury". They both showed declines in linguistic complexity while they are under 30 years old. Because lexical complexity can be preserved during normal aging (Burke & Shafto, 2008), the decline is very rare in their ages. Veterans such as Tom Brady who does not have official records of on field concussion, also showed declines in the linguistic complexity. In summary, this study found linguistic decline in players who had concussion, as well as players who started football careers in NFL years ago. I do not intend to draw a strong conclusion of the relationships between the history of playing football, record of on field concussion and decline in linguistic complexity, due to the small sample size. However, we cannot ignore the fact that very few other factors such as age and education can explain the significant differences in the study.

This study is novel. To my knowledge, this data on NFL is the first set of data that aim to track linguistic changes to reflect the status of NFL players' cognitive-linguistic health. It is very hard to draw causal conclusion between linguistic changes and MTBI. However, in contrast with the group of non-player executives who never played football, these players with declines in lexical complexity showed that they are at least at risk of CTE. These findings support the claim that CTE may be linked to history of playing football, especially in professional football leagues such as NFL (Omalu et al., 2010).

Effects of Age on Changes in Linguistic Complexity

Linguistic complexity changes as a result of aging. Normal aging can be also related to decreasing linguistic complexity (Burke & Shafto, 2008), however researchers were able to separate the effects of normal aging and dementia to some extent. Kemper et al. (1989) reported that the young adult group (18 to 28 years old) produced more sentence segments (number of clauses) than older adult groups (i.e., 60+, 70+, and 80+) on language samples from oral and written statements. This indicates that aging relates to the loss of syntactic complexity due to the shrinkage of working memory (Burke & Shafto, 2008). Because processing the structure of a sentence is from left to right. Thus the left most part of a sentence has to stay in the working memory while the remaining right part of a sentence is about to say (Burke & Shafto, 2008). Reduced working memory leads to reduced ability of processing syntactic complexity. Kemper et al. (2001) reported that grammatical and semantic content complexity did not decline rapidly until mid 70s in the normal aging group, while complexity declined rapidly regardless of the subject's age in the group of dementia patients. Reviews by Burke & Shafto (2008) showed that semantic complexity such as vocabulary size is well preserved in the normal aging group, as compared to younger group. The ability of syntactic information processing is constrained in normal aging group, as compared to younger group due to shrinking capability of working memory (Burke & Shafto, 2008). Le et al. (2011)

summarized several studies that compared semantic and syntactic complexity in the normal aging group and groups with dementia. They concluded that dementia accelerates the degradation in measures of both lexical complexity and semantic complexity while normal aging group can preserve some level of lexical complexity. The results of this NFL study support that age does not play a role in the changes of linguistic patterns. On average, non-player executives are older than the players I analyzed; however, their linguistic complexity showed little decline (see Table 6 and Table 7).

Subtle yet Consistent Changes in Linguistic Complexity over Time

For adults' post education, linguistic complexity patterns are very stable (Burke & Shafto, 2008); both case study I and case study II support this finding. To discuss changes of linguistic patterns in my paper, I define a term called rate of change, which is the Δ TTR (or Δ LD) / Duration of interview range for the corresponding person. It roughly reflects how fast the linguistic patterns change. Rates of non-player and player groups are put together to give a comprehensive view of their linguistic patterns, as shown in Figure 6. Mean rates of 27 counts of significant trends is -0.097% with SEM = 0.225%. Each dot represents a significant (increasing or decreasing) trend. For comparison between two studies, Reagan's TTR rate (0.276%) is also plotted as a reference line. The patterns I found between the non-player group and the player group are more distinctive when rates of changes in linguistic complexity considered. These rates from results of NFL study are very mild, as compared to Reagan's data. Because MTBI and CTE can lead to serious disorders such as dementia later in life (Guskiewicz et al., 2005), these mild decline in linguistic complexity at young athletes need more attention. For example, tracking language changes in football players regularly help monitor the cognitive health status. Warning early signs of cognitive declines in such measures help additional intervention.

Limitations

Limited power of correlation. It is hard to draw causality from simple linear regression analysis. Injuries to the brain such as MTBI and neurological disorders such as AD are not the only factors that lead to observations in changes of linguistic complexity over time. The power of correlation is further limited when data sources are not controlled in advance. Although my results cannot draw causality between MTBI and declines in linguistic complexity in NFL players, the results from my studies are consistent with the findings of others. Studies have shown that the existence of TBI can adversely affect a broad range of cognitive-linguistic functions including naming and discourse tasks (King et al. 2006; Tucker and Hanlon, 2009). The findings on early decline of linguistic complexity in those players is an indicator of the adverse effect of repeated head impacts they receive during the play.

Limitation of statistical significance power. Those statistical significant results should be interpreted with caution. In the present longitudinal study, the length of data varies from 1 year to 7 years long (see Table 4 and Table 5). This makes the criterion for each linear regression varies based on their own length of data because of the nature of linear regression analysis.

Limited numbers of metrics. In the present study, I only adapted two metrics from previous studies (Le et al., 2011, Snowden et al., 1996, Kemper et al., 1989. Kemper et al., 2001, Hier et al., 1985). These two metrics link with the onset and development of dementia. It is certainly possible that composite or weighted scores of several standard measures give us more power beyond standalone correlation analysis within each metric. Advanced techniques, such as models from Machine Learning, may be useful in examining longitudinal linguistic data.

Future Directions

Further cleaning in existing NFL data. Currently, data sets in the NFL study use all all available transcripts with desired formats from both players and non-players. However, news conferences often include sub-categories like pre-game news conferences, post-game new conferences and off-season interviews. The current case study doesn't separate and differentiate transcribed answers from these categories especially for players. Further analysis should test whether language ability of players are the same in performance of language abilities in different situations, i.e., the before-game stage and the after-game stage.

New methods. Further analysis may include syntactic complexity measures. Researchers have suggested that the progression of dementia causes difficulties in lexical access which might lead to increased use of empty nouns and filler words (Kemper et al., 2001; Hier et al., 1985; MacDonald et al., 2001). However, it is not clear that how to distinguish normal aging and AD type dementia, if higher-level information such as syntactic complexity is included in the analysis (Le et al., 2011). It may be difficult to distinguish syntactic complexity between normal aging and dementia group, because both groups can show decline in syntactic complexity (Cheung & Kemper, 1992; Kemper et al., 2001; Burke & Shafto, 2008). Careful selection of measures of syntactic complexity is certainly crucial. Because no conclusion could be drawn using either mean length per utterance or Yngve depth (Yngve 1960) syntactic complexity (Le et al., 2011).

Furthermore, new smoothing methods are also required in syntactic complexity measure. Currently, a 1000 words sliding window removes punctuations that are needed for sentence level syntactic complexity analysis. Smoothing methods that preserve both intact structure of sentences and compatibility of word level lexical complexity analysis will be helpful.

New Data. Besides digging data from web pages, new data can come from people's daily activities. I analyzed existing public data to show the possibility of tracking changes in linguistic patterns of transcribed speech samples. These case studies showed that it is possible to use speech and language data to track cognitive-linguistic decline. Future work could incorporate background phone applications to monitor complexity in the speech from talk or text messages. Consistent decline in linguistic complexity may indicate a warning sign of the onset of dementia. Early interventions may further help people at risk to slow down or prevent the development of disorders.

REFERENCES

Alzheimer’s Association. (2014). 2014 Alzheimer’s disease facts and figures. *Alzheimer’s & Dementia, 10*(2), e47-e92.

Bailes, J. E., Petraglia, A. L., Omalu, B. I., Nauman, E., & Talavage, T. (2013). Role of subconcussion in repetitive mild traumatic brain injury: a review. *Journal of neurosurgery*, *119*(5), 1235-1245.

Bates, E., Harris, C., Marchman, V., Wulfeck, B., & Kritchevsky, M. (1995). Production of complex syntax in normal ageing and Alzheimer’s disease. *Language and Cognitive Processes, 10*(5), 487–539.

Beckwith, J. G., Greenwald, R. M., Chu, J. J., Crisco, J. J., Rowson, S., Duma, S. M., … others. (2013). Timing of concussion diagnosis is related to head impact exposure prior to injury. *Medicine and Science in Sports and Exercise, 45*(4), 747-754.

Benson, B. W., McIntosh, A. S., Maddocks, D., Herring, S. A., Raftery, M., & Dvořák, J. (2013). What are the most effective risk-reduction strategies in sport concussion? *British Journal of Sports Medicine, 47*(5), 321-326.

Berisha, V., Wang, S., LaCross, A., & Liss, J. (2015). Tracking discourse complexity preceding Alzheimer’s disease diagnosis: a case study comparing the press conferences of presidents Ronald Reagan and George Herbert walker bush. *Journal of Alzheimer’s Disease: JAD, 45*(3), 959-963.

Bird, H., Ralph, M. A. L., Patterson, K., & Hodges, J. R. (2000). The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and Language, 73*(1), 17-49.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.

Blazer, D. G., & Steffens, D. C. (2009). *The American Psychiatric Publishing Textbook of Geriatric Psychiatry*. 4th edition. Arlington, VA: American Psychiatric.

Borgaro, S. R., Prigatano, G. P., Kwasnica, C., & Rexer, J. L. (2003). Cognitive and affective sequelae in complicated and uncomplicated mild traumatic brain injury. *Brain Injury, 17*(3), 189-198.

Botel, M., & Granowsky, A. (1972). A formula for measuring syntactic complexity: A directional effort. *Elementary English, 49*(4), 513-516.

Burke, D. M., Shafto, M. A. (2008). Language and aging. *The Handbook of Aging and Cognition, 3*, 373-443.

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics, 13*(01), 53–76.

Christopher, M. B. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics): Springer-Verlag New York, Inc.

Cobb, B. R., Urban, J. E., Davenport, E. M., Rowson, S., Duma, S. M., Maldjian, J. A., ... & Stitzel, J. D. (2013). Head impact exposure in youth football: elementary school ages 9–12 years and the effect of practice structure. *Annals of biomedical engineering*, *41*(12), 2463-2473.

Cook, C., Fay, S., & Rockwood, K. (2009). Verbal repetition in people with mild-to-moderate Alzheimer disease: a descriptive analysis from the VISTA Clinical Trial. *Alzheimer Disease & Associated Disorders, 23*(2), 146–151.

Cummings, J. L., Darkins, A., Mendez, M., Hill, M. A., & Benson, D. F. (1988). Alzheimer's disease and Parkinson's disease Comparison of speech and language alterations. *Neurology, 38*(5), 680–680.

Daniel, R. W., Rowson, S., & Duma, S. M. (2012). Head impact exposure in youth football. *Annals of Biomedical Engineering, 40*(4), 976–981.

Davis, G. A., O'Neil-Pirozzi, T. M., & Coon, M. (1997). Referential cohesion and logical coherence of narration after right hemisphere stroke. *Brain and Language, 56*(2), 183–210.

Dodge, H. H., Mattek, N., Gregor, M., Bowman, M., Seelye, A., Ybarra, O., ... & Kaye, J. A. (2014). Social Markers of Mild Cognitive Impairment: Proportion of Word Counts in Free Conversational Speech. *Current Alzheimer Research, 12*(6), 513-519.

Duma, S. M., & Rowson, S. (2014). The biomechanics of concussion: 60 years of experimental research. *Concussions in Athletics*, 115–137.

Fazio, V. C., Lovell, M. R., Pardini, J. E., & Collins, M. W. (2007). The relation between post concussion symptoms and neurocognitive performance in concussed athletes. *NeuroRehabilitation-An Interdisciplinary Journal*, *22*(3), 207-216.

Frazier, Lyn. "Syntactic complexity", Natural language parsing. Ed.David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky.. 1st ed. Cambridge: Cambridge University Press, 1985. 129-189. Cambridge Books Online.

Gottschalk, L. A., Uliana, R., & Gilbert, R. (1988). Presidential candidates and cognitive impairment measured from behavior in campaign debates. *Public Administration Review*, 613-619.

Guskiewicz, K. M., Marshall, S. W., Bailes, J., McCrea, M., Cantu, R. C., Randolph, C., & Jordan, B. D. (2005). Association between recurrent concussion and late-life cognitive impairment in retired professional football players. *Neurosurgery, 57*(4), 719-726.

Hier, D. B., Hagenlocker, K., & Shindler, A. G. (1985). Language disintegration in dementia: Effects of etiology and severity. *Brain and Language, 25*(1), 117–133.

Holm, H., Mignéus, M., & Ahlsén, E. (1994). Linguistic symptoms in dementia of Alzheimer type and their relation to linguistic symptoms of aphasia. *Scandinavian Journal of Logopedics and Phoniatrics, 19*(3), 99–106.

Huberman, B. A., & Adamic, L. A. (1999). Internet: growth dynamics of the world-wide web. *Nature*, *401*(6749), 131-131.

Kemper, S., Kynette, D., Rash, S., O'Brien, K., & Sprott, R. (1989). Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics, 10*(01), 49–66.

Kemper, S., Thompson, M., & Marquis, J. (2001). Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging, 16*(4), 600.

King, K. A., Hough, M. S., Walker, M. M., Rastatter, M., & Holbert, D. (2006). Mild traumatic brain injury: Effects on naming in word retrieval and discourse. *Brain Injury, 20*(7), 725–732.

Kinsella, G. J., Mullaly, E., Rand, E., Ong, B., Burton, C., Price, S., ... & Storey, E. (2009). Early intervention for mild cognitive impairment: a randomised controlled trial. *Journal of Neurology, Neurosurgery & Psychiatry, 80*(7), 730-736.

Iaccarino, L., Crespi, C., Della Rosa, P. A., Catricalà, E., Guidi, L., Marcone, A., ... & Perani, D. (2015). The semantic variant of primary progressive aphasia: clinical and neuroimaging evidence in single subjects. *PloS one, 10*(3), e0120197.

Lee, L. L. (1974). Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians. Northwestern University Press.

Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, *26*(4).

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36.

MacDonald, M. C., Almor, A., Henderson, V. W., Kempler, D., & Andersen, E. S. (2001). Assessing working memory and language comprehension in Alzheimer's disease. *Brain and language, 78*(1), 17-42.

Manly, J. J., Schupf, N., Tang, M. X., & Stern, Y. (2005). Cognitive decline and literacy among ethnically diverse elders. *Journal of Geriatric Psychiatry and Neurology, 18*(4), 213-217.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1-86.

McKee, A. C., Daneshvar, D. H., Alvarez, V. E., & Stein, T. D. (2014). The neuropathology of sport. *Acta neuropathologica*, *127*(1), 29-51.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men?. *Science, 317*(5834), 82-82.

Mez, J., Stern, R. A., & McKee, A. C. (2013). Chronic traumatic encephalopathy: Where are we and where are we going? *Current Neurology and Neuroscience Reports, 13*(12), 1–12.

Nicholas, M., Obler, L. K., Albert, M. L., & Helm-Estabrooks, N. (1985). Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech, Language, and Hearing Research, 28*(3), 405–410.

Nordström, P., Michaëlsson, K., Gustafson, Y., & Nordström, A. (2014). Traumatic brain injury and young onset dementia: a nationwide cohort study. *Annals of neurology*, *75*(3), 374-381.

Nowinski, C. (2006). *Head games: football's concussion crisis from the NFL to youth lea Omalu, B. I., DeKosky, S. T., Minster, R. L., Kamboh, M. I., Hamilton, R. L., & Wecht, C. H. (2005). Chronic traumatic encephalopathy in a National Football League player. Neurosurgery, 57(1), 128-134*.

Omalu, B. I., DeKosky, S. T., Minster, R. L., Kamboh, M. I., Hamilton, R. L., & Wecht, C. H. (2005). Chronic traumatic encephalopathy in a National Football League player. *Neurosurgery, 57*(1), 128-134.

Omalu, B. I., Hamilton, R. L., Kamboh, M. I., DeKosky, S. T., & Bailes, J. (2010). Chronic traumatic encephalopathy (CTE) in a National Football League Player: Case report and emerging medicolegal practice questions. *Journal of forensic nursing, 6*(1), 40-46.

Padrez, K. A., Ungar, L., Schwartz, H. A., Smith, R. J., Hill, S., Antanavicius, T., … Merchant, R. M. (2015). Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ Quality & Safety*, bmjqs–2015.

Pellman, E. J., Viano, D. C., Casson, I. R., Tucker, A. M., Waeckerle, J. F., Powell, J. W., & Feuer, H. (2004). Concussion in professional football: Repeat injuries—Part 4. *Neurosurgery*, *55*(4), 860-876.

Poole, V. N., Breedlove, E. L., Shenk, T. E., Abbas, K., Robinson, M. E., Leverenz, L. J., … Talavage, T. M. (2015). Sub-concussive hit characteristics predict deviant brain metabolism in football athletes. *Developmental Neuropsychology, 40*(1), 12–17.

Roberts, R. O., & Petersen, R. C. (2014). Predictors of early-onset cognitive impairment. *Brain, 137*(5), 1280-1281.

Rosenberg, S., & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics, 8*(01), 19-32.

Schwarz, A. (2010). Suicide reveals signs of a disease seen in NFL. New York Times.

Smith, S. R., Chenery, H. J., & Murdoch, B. E. (1989). Semantic abilities in dementia of the Alzheimer type. II. Grammatical semantics. *Brain and Language*, *36*(4), 533-542.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (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.

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J., & Lewkowicz, D. J. (2012). The development of audiovisual speech perception. *Multisensory Dev*, 207-228.

Stamford, J. A., Schmidt, P. N., & Friedl, K. E. (2015). What Engineering Technology Could Do for Quality of Life in Parkinson's Disease: A Review of Current Needs and Opportunities. *Biomedical and Health Informatics, IEEE Journal of, 19*(6), 1862-1872.

Tucker, F. M., & Hanlon, R. E. (1998). Effects of mild traumatic brain injury on narrative discourse production. *Brain injury, 12*(9), 783-792.

Wechsler, D. (2008). *Wechsler adult intelligence scale-fourth*. San Antonio: Pearson.

Yngve, V. H. (1960). A model and an hypothesis for language structure. Proceedings of the American philosophical society, 104(5), 444-466.

Table 1

Counting Statistics from Reagan and Bush's Data

	Ronald Reagan Mean (SD)	George H.W. Bush Mean (SD)
Unique words	413.6 (17.2)	405.4 (17.6)
Non-specific nouns + filler words	25.1 (6.3)	21.5 (4.3)
low-image verbs	94.6 (10.0)	110.5 (11.9)

Note. This table is derived from Berisha et al., 2015 with permission.

Table 2

Summary of Linear Regression Results

	Ronald Reagan		George H.W. Bush	
	r-value	p-value	r-value	p-value
Unique words	**-0.446**	**0.002**	-0.098	0.343
Non-specific nouns + filler words	**0.358**	**0.017**	0.053	0.608
Low-image verbs	0.032	0.835	-0.099	0.333

Note. This table is derived from Berisha et al., 2015 with permission. Bold values are significant.

Table 3

Available Teams and Sample Links

Team Name	Sample link
New England Patriots	http://www.patriots.com/news/2016/01/14/bill-belichick-press-conference-transcript-114
New York Giants	http://www.giants.com/news-and-blogs/article-1/Quotes-129-Coach-Tom-Coughlin/0e263c65-d49a-403e-8d1d-58bdc2284f08
Washington Redskins	http://www.redskins.com/news-and-events/article-1/Quotes-Jay-Gruden-011116/0debbf03-a289-47d6-94d6-756d6c0834c6
New York Jets	http://www.newyorkjets.com/news/article-5/REX-It-Can-Get-Ugly-Out-There-During-OTAs/a78a19f3-e526-48e2-855d-34ae0a37391f
Atlanta Falcons	http://www.atlantafalcons.com/news/article-1/Transcripts-Smith-Ryan-Press-Conferences/b6c04644-5500-4835-85f0-1361d6a17e44
Seattle Seahawks	http://seahawksmedia.seainternet.com/Transcripts/Carroll.htm
Baltimore Ravens	http://www.baltimoreravens.com/news/transcripts.html#cltop_1454255525511
Cincinnati Bengals	http://www.bengals.com/news/article-1/Marvin-Lewis-News-Conference-Transcript/aecbf930-01c9-4560-9887-e8e0dc425ee0
Green Bay Packers	http://www.packers.com/news-and-events/article-1/Mike-McCarthy-Press-Conf-Transcript---Jan-20/c202e922-ea4e-4f6a-8360-23cc70facdd0
Houston Texans	http://media.houstontexans.com/section_display.cfm?section_id=259
Pittsburgh Steelers	http://www.steelers.com/news/article-1/Transcript-Sammie-Coates-conference-call/b2f7a8b0-414c-41cd-82b8-76216af05bf1

Table 4

Summary of Demographical Information on Players

Name	Interview range		Age	Education	Initial LD	Initial TTR
Tom Brady	Nov-07	Feb-15	38	4	0.264	0.455
Vince Wilfork	Jun-08	Jan-14	34	3	0.263	0.439
Eli Manning	Feb-10	Jun-15	35	4	0.285	0.434
Russell Wilson	Jan-13	Jun-15	27	4	0.254	0.459
Richard Sherman	Jul-13	Jun-15	27	4	0.266	0.442
Ander Johnson	Feb-08	May-15	34	3	0.232	0.454
Andy Dalton	Mar-12	Jun-15	28	4	0.258	0.448
Carson Palmer	Mar-10	Jun-11	36	4	0.269	0.452
Mark Sanchez	Feb-11	Oct-13	29	3	0.283	0.452
RG III	Nov-13	Jun-14	25	3	0.267	0.447

Table 5

Summary of Demographic Information on Non-player Executives

Name	Interview dates range		Age	Education	Initial LD	Initial TTR
Bill Belichick	Nov-07	Feb-15	63	4	0.264	0.448
Dean Pees	Feb-08	Jun-10	66	4	0.246	0.435
Nick Caserio	Jan-10	Jan-15	40	4	0.268	0.425
Matt Patricia	Apr-11	Jul-14	41	4	0.220	0.448
Josh McDaniels	Jan-08	Jul-14	39	3	0.259	0.437
Tom Coughlin	Feb-10	Jun-15	69	4	0.285	0.434
Pete Carroll	Jan-13	Jun-15	64	4	0.269	0.435
Marvin Lewis	Apr-09	Jun-15	57	?	0.280	0.459
Rex Ryan	Feb-11	Nov-14	53	6(?)	0.278	0.443
Mike McCarthy	Mar-07	Mar-12	52	6	0.292	0.453
Mike Tomlin	Jan-09	Feb-15	43	4	0.292	0.433
Wade Phillips	Jul-11	Jun-14	68	2	0.253	0.443
Cam Cameron	Nov-08	May-13	54	4	0.271	0.446
Jerry Rosburg	Mar-09	Jan-15	60	4	0.279	0.444
John Harbaugh	Jan-09	Jun-15	53	4	0.271	0.447
Mike Shanahan	Oct-11	Jun-14	63	5	0.264	0.445
Bill Obrien	Jun-14	Jun-15	46	2	0.280	0.436
Rick Smith	Oct-07	Aug-14	40	4	0.281	0.443

Note. "?" means no data available or can not confirm the data source.

Table 6

Summary of Linear Regression Analysis between Time Stamps and LD on Non-player Executives and Players

Non-player	*r*-value	*p*-value	ΔLD	Player	r-value	p-value	ΔLD
Bill Belichick	0.073	0.003	0.004	Tom Brady	-0.116	0.022	-0.006
Dean Pees	0.430	0.002	0.021	Vince Wilfork	0.072	0.610	
Nick Caserio	0.475	<0.0001	0.028	Eli Manning	0.266	0.002	0.013
Matt Patricia	0.245	0.083		Russell Wilson	-0.069	0.350	
Josh McDaniels	0.379	0.000	0.018	Richard Sherman	0.217	0.078	
Tom Coughlin	0.266	0.002	0.013	Andre Johnson	-0.287	0.022	-0.014
Pete Carroll	0.141	0.001	0.007	Andy Dalton	-0.305	0.004	-0.015
Marvin Lewis	-0.260	0.000	-0.013	Carson Palmer	0.063	0.632	
Rex Ryan	-0.140	0.004	-0.008	Mark Sanchez	0.050	0.634	
Mike McCarthy	0.171	<0.0001	0.009	RG III	-0.263	0.054	
Mike Tomlin	0.250	0.037	0.034				
Wade Phillips	0.359	0.011	0.015				
Cam Cameron	0.175	0.078					
Jerry Rosburg	-0.071	0.644					
John Harbaugh	-0.103	0.075					
Mike Shanahan	0.082	0.321					
Bill Obrien	0.323	0.018	0.023				
Rick Smith	-0.038	0.750					

Note. Data which showed significant increasing trends are marked in green, while those showed significant decreasing trend are marked in red.

Table 7

Summary of Linear Regression Analysis between Time Stamps and TTRs on Non-Player Executives and Players

Non-player	r-value	p-value	ΔTTR	Player	r-value	p-value	ΔTTR
Bill Belichick	0.029	0.229		Tom Brady	-0.211	<0.0001	-0.013
Dean Pees	0.036	0.800		Vince Wilfork	-0.469	<0.0001	-0.028
Nick Caserio	0.107	0.281		Eli Manning	-0.080	0.349	
Matt Patricia	0.549	<0.0001	0.044	Russell Wilson	0.068	0.358	
Josh McDaniels	0.213	0.032	0.007	Richard Sherman	0.088	0.477	
Tom Coughlin	-0.080	0.349		Andre Johnson	0.012	0.925	
Pete Carroll	-0.092	0.030	-0.004	Andy Dalton	-0.426	<0.0001	-0.036
Marvin Lewis	-0.060	0.267		Carson Palmer	0.048	0.715	
Rex Ryan	-0.138	0.004	-0.009	Mark Sanchez	-0.321	0.002	-0.019
Mike McCarthy	0.076	0.040	0.003	RG III	-0.534	<0.0001	-0.028
Mike Tomlin	0.097	0.422					
Wade Phillips	-0.029	0.841					
Cam Cameron	0.003	0.978					
Jerry Rosburg	-0.293	0.050					
John Harbaugh	0.105	0.070					
Mike Shanahan	-0.003	0.970					
Bill Obrien	-0.570	<0.0001	-0.029				
Rick Smith	-0.104	0.378					

Note. Data which showed significant increasing trends are marked in green, while those showed significant decreasing trend are marked in red.

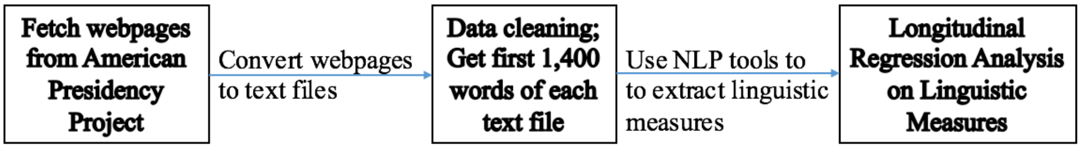

Figure 1. General Methodology for Getting and Cleaning the Data.

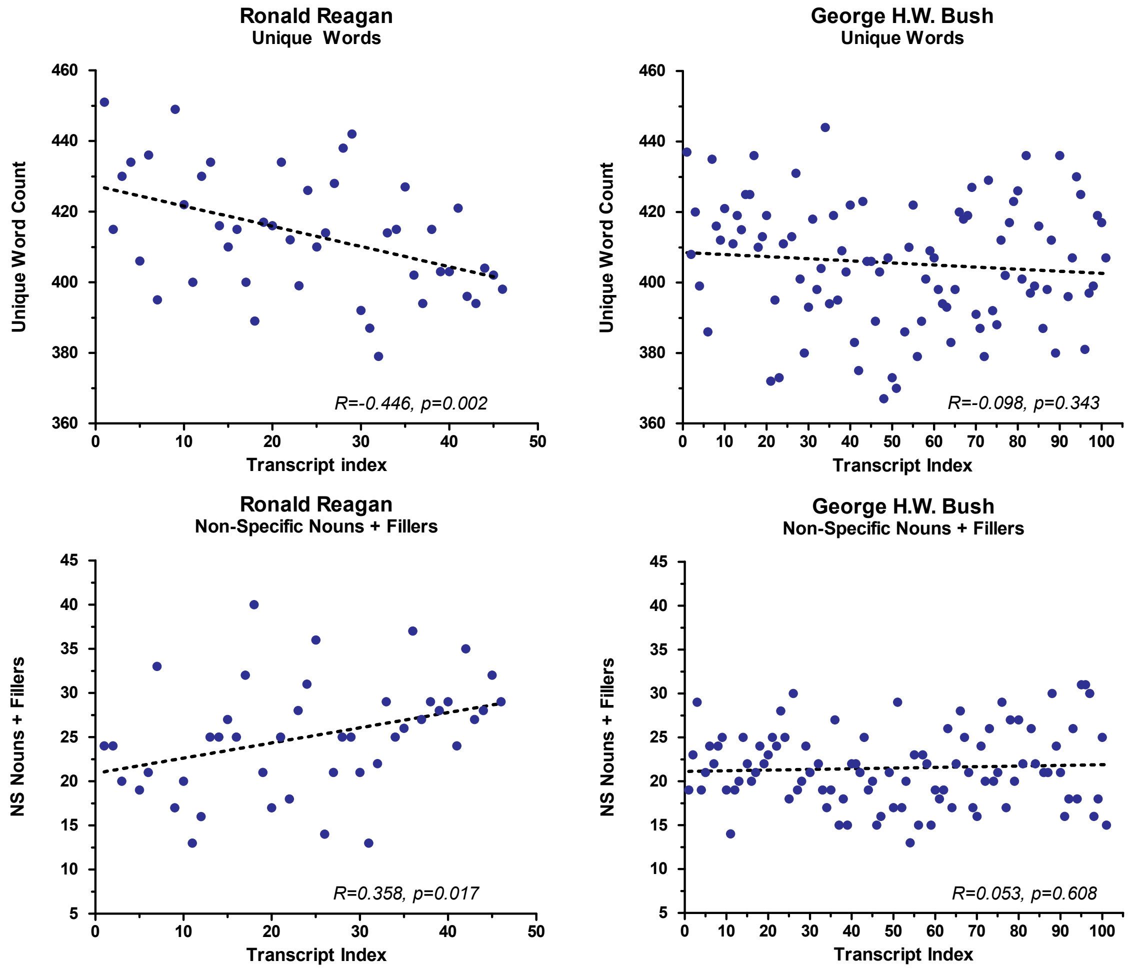

Figure 2. Pairwise Comparsion between Reagan and Bush in Trends of Lingustic Complexity. Left panel: Reagan's count of unique words and count of non-specific Nouns+fillers vs transcript index. Right panel: Bush's count of unique words and count of non-specific Nouns+fillers vs transcript index. The figure is derived from Berisha et al., 2015 with permission.

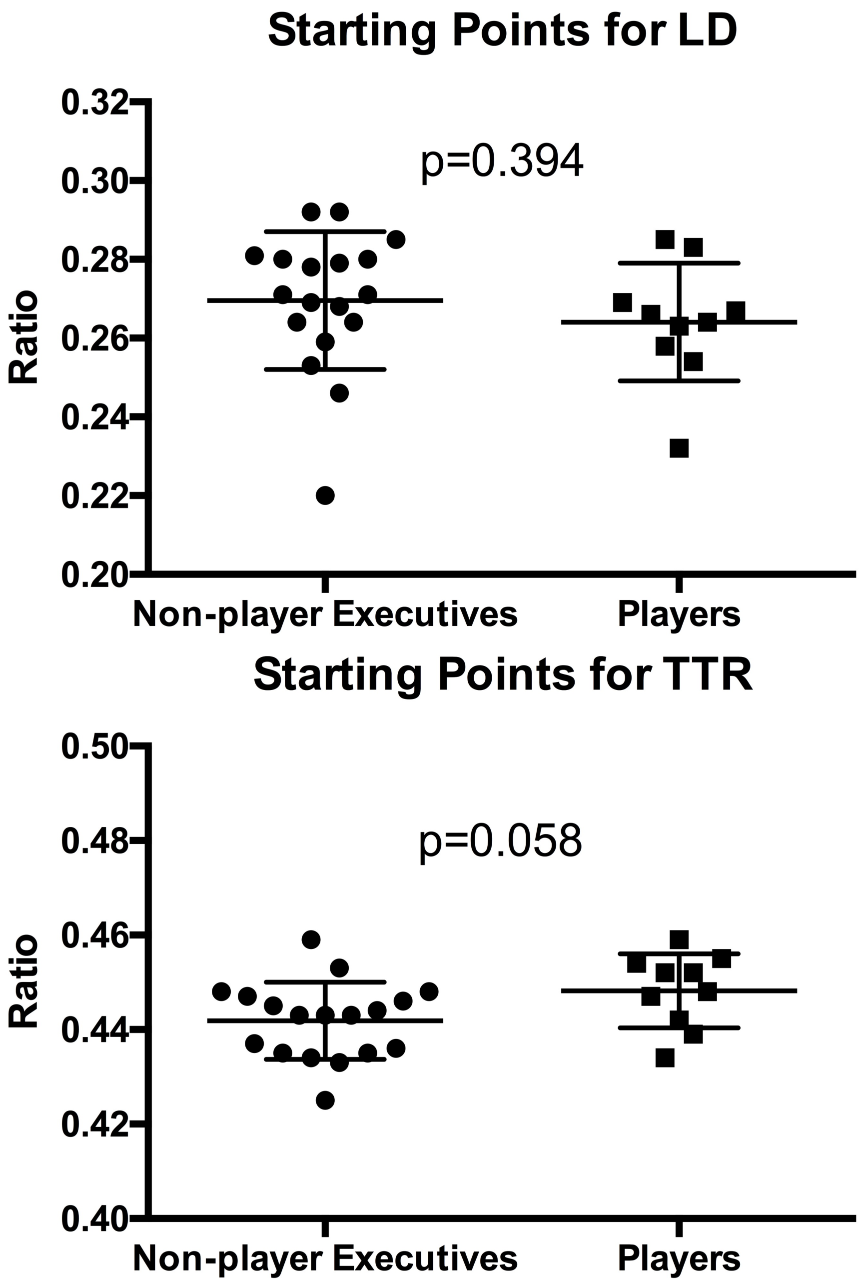

Figure 3. Estimated Starting Points for LDs and TTRs in Each Group.

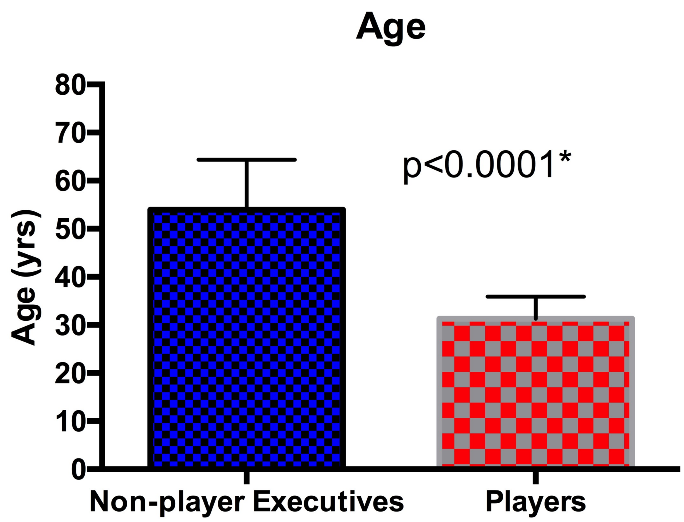

Figure 4. Comparison of Non-player Executives' Current Ages and Players' Current Ages.

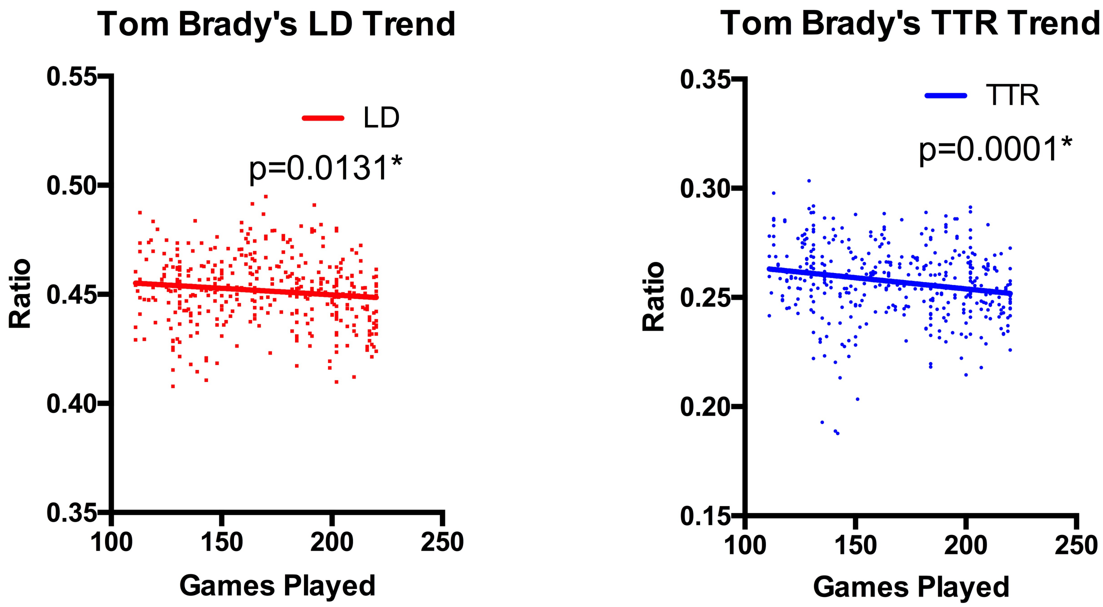

Figure 5. Plot of Tom Brady's LDs and TTRs as a Function of Number of Games Played.

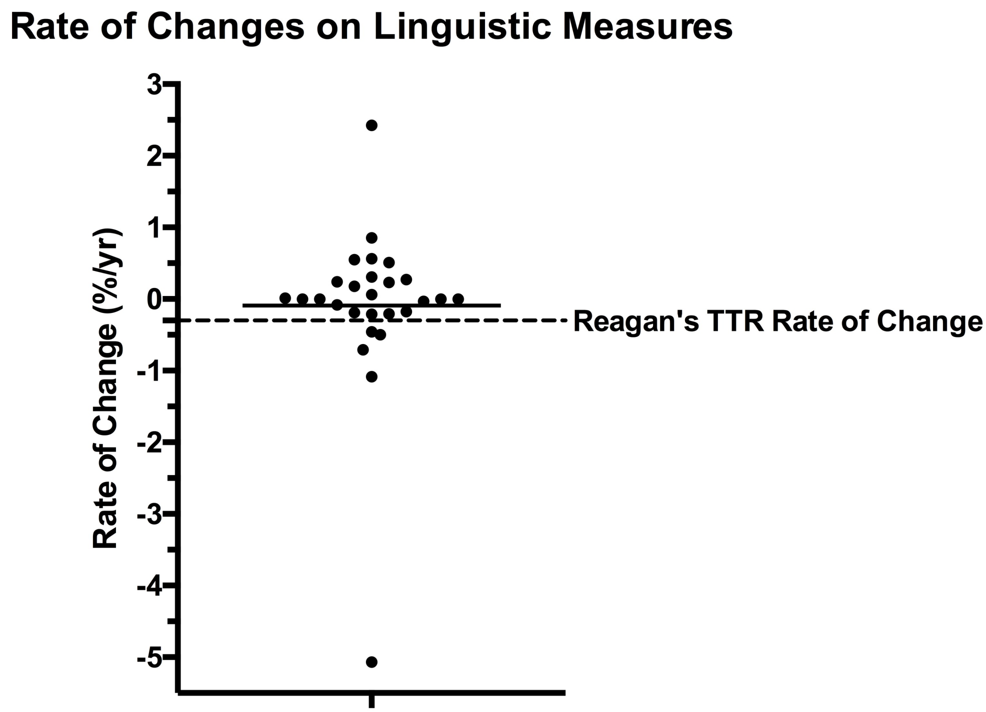

Figure 6. Summary of Rate of Changes from Significant Linear Regression Results in the Linguistic Complexity Longitudinal Analysis.