AUTOMATIC TEXT AND SPEECH PROCESSING FOR THE DETECTION OF DEMENTIA

by

Kathleen Fraser

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Automatic text and speech processing for the detection of dementia

Kathleen Fraser

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

2016

Dementia is a gradual cognitive decline that typically occurs as a consequence of neurodegenerative disease, and can result in language deficits (i.e. aphasia). I show that linguistic features automatically extracted from the connected speech samples of individuals with dementia can both differentiate these individuals from healthy controls and contribute to our knowledge of the nature of language impairment in dementia. As a secondary goal, I address the challenges of a fully automated processing pipeline.

I focus on a dementia syndrome known as primary progressive aphasia (PPA), in which language abilities are specifically impaired. I begin by automatically extracting linguistic information from transcripts of PPA speech, training machine learning classifiers to differentiate between the different variants of PPA relative to healthy controls, and interpreting the selected features in the context of the PPA literature. While traditional measures of syntactic complexity do not distinguish between the groups, the inclusion of parse-based syntactic features ultimately leads to accuracies of over 90% in three classification tasks.

Having shown that the extracted features can differentiate the groups, I examine how these features degrade as a result of the processing steps in a fully automated pipeline, including automatic speech recognition (ASR) and sentence segmentation. The classifiers still achieve positive results, although the degraded feature accuracy may be of concern in biomedical applications.

I then explore a question of some debate in the literature: Is there a difference between agrammatism in PPA and agrammatism in post-stroke aphasia? Above-baseline classification results suggest that there are indeed differences between these two impairments.

Having validated the methodology on PPA, I conclude by examining whether a similar analysis will detect Alzheimer's disease from speech samples, even though language impairment is not the primary symptom of the disease. By including additional features to measure the information content of the narrative, classification accuracies of up to 81% are achieved. I repeat the classification experiment using ASR transcripts, and find that many of the relevant features are still significantly different between the groups, suggesting that a fully automated analysis may be possible as ASR improves.

Dedicated to Elda May Hepburn Fraser (1915–2014)

You continue to inspire.

Acknowledgements

I wish to gratefully acknowledge funding from an NSERC CGS-D scholarship, as well as grants to my collaborators from NSERC, CIHR, and the Alzheimer's Association. I would also like to express my deepest gratitude to all the volunteers who contributed speech samples to the data sets discussed herein, without whom this research would not be possible.

Thank you to my supervisors, for their patience and guidance. To Graeme, and whatever inscrutable metric you use to select graduate students, for giving me this opportunity. Thank you for pushing me, for keeping me on track despite all the distractions I generated for myself, and for only making fun of my accent every once in a while. Thank you to Jed, for your unfailing enthusiasm and optimism, and for always reminding me of the "big picture" perspective when I needed it the most.

To my other committee members: Elizabeth, you shared your data with me and made it seem from my perspective like the easiest thing in the world. I realize now that it was complicated and exceptionally generous — thank you. Frank, I still have no idea how you accomplish so much in the same 24 hours in which I manage to do so little; thank you for always making time for me. Thank you also to my external examiners, Brian Roark and Regina Jokel, for your inspiring work in this area, and your thought-provoking questions and suggestions for my research.

To my other collaborators: It has been a pleasure to learn from you, and I hope we can continue many of these collaborations into the future. Thank you for sharing your expertise, time, data, and in some cases even your homes: Naida Graham, Sandra Black, Carol Leonard, Karine Marcotte, Cynthia Thompson, Jennifer Mack, Martin van Schijndel, Sandra Maria Alusio (and her students), and Leticia Mansur. An especially big thank you to Naama Ben-David and Luke Zhou, for working so hard over the summer and then still answering my emails long after the summer was over. I know you will be successful at whatever you choose to do.

Thank you to all my fellow graduate students for your advice and support, especially my

office mates (at one time or another) Aditya, Tong, Vanessa F., and Maria, my Pratt lunch buddies Varada, Nona, Jackie, Aida, and Afsaneh, and my best Brazilian friend Vanessa M. A special thank you to Libby, partly because I know you read these things, but really for your help guiding me through matters both professional and personal. I am grateful that our paths happened to overlap exactly as and when they did. Cheers also to the rest of the CL group for many interesting and funny conversations over the years.

To my family, for really not being into computers at all, for fostering my interests in art and language, and for raising their eyebrows only ever so slightly when I chose this road instead, and loving me all the same. In particular to my mom, for never hesitating to do whatever needed to be done to help me succeed, despite your own busy schedule. You made the impossible possible on a regular basis. Thank you.

Finally, to my husband Danny, I never could have done it without you — or at least, I wouldn't have wanted to. You kept the lights on and the home fires burning, all while working full time in your own career and being an equal partner in every way. I love you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Dementia is a degenerative cognitive impairment, with a severity that interferes with the ability to function normally in day-to-day life. Prince et al. (2013) estimate that 35.6 million people worldwide were living with dementia in 2010, and the Alzheimer Society of Canada[1] reports that 747,000 Canadians suffered from cognitive impairment or dementia in 2011. These numbers are projected to double in the next 20 years (Prince et al., 2013). The increasing prevalence of dementia is expected to lead to more pressure on the healthcare system to provide more services related to dementia. Such services could include screening for early detection of cognitive decline, accurate diagnosis of different types of dementia, and long-term care of people with dementia, including assistive technologies to allow individuals to safely remain in their homes and communities. Tools for monitoring the severity of dementia symptoms, and tracking the efficacy of potential interventions, will also be required.

A diagnosis of dementia is made on the basis of several factors, including neuropsychological testing, interviews with family and caregivers, and brain imaging. In some types of dementia, such as primary progressive aphasia (which will be the main focus of this dissertation), language deficit is a core symptom. Language ability is assessed by a range of tests, including single-word naming, recitation of overlearned sequences (such as counting, or days of the week), repetition of words and sentences, fluency tests (such as naming animals), and

[1] `http://www.alzheimer.ca`

many others. These tests are designed to examine specific language processes, allowing a clinician to determine a precise pattern of language ability and disability. However, these tests do not necessarily reflect how people use language in everyday life, which requires the production of complete sentences, expressing a range of meaning and emotion. Individuals who perform well on single-word naming tasks may not be able to string those words together to form grammatical utterances. Or the opposite can be true: patients who perform poorly on very specific semantic tasks could still be able to produce quite long and fluent sentences.

Many researchers and clinicians therefore agree that the analysis of narrative or conversational speech is important for assessing the extent of an individual's language impairment (Bucks et al., 2000; Forbes-McKay and Venneri, 2005; Rohrer et al., 2008; Sajjadi et al., 2012). However, the prevailing view is summarized by Sajjadi et al. (2012) in their paper on language in dementia: "Connected speech provides the most realistic measure of language function but its use has been restricted by operational constraints." There is simply not enough time to perform a quantitative analysis of long streams of speech in a clinical setting. Some researchers have approached this problem by using methods from natural language processing (NLP) to automate the analysis of speech in various neurodegenerative disorders (for example, Thomas et al. (2005); Jarrold et al. (2010); Roark et al. (2011)). The basic processing pipeline for these studies (and the work presented here) is shown in Figure 1.1, although not all studies use each component in the diagram. These studies have shown that it is possible for machine learning classifiers to achieve high accuracy on some diagnostic tasks, when trained on features which were automatically extracted from speech transcripts.

However, previous computational work suffers from the following limitations. In general, there is an over-emphasis on classification accuracy, and a lack of attention paid to the features that form the foundation for the classification results. For example, Peintner et al. (2008) only report the number of statistically significant features, rather than listing the features themselves. Jarrold et al. (2014) discuss only 8 of the apparently hundreds of features they extract (indeed, the exact number and nature of the features is not clear). Roark et al. (2011) present the values

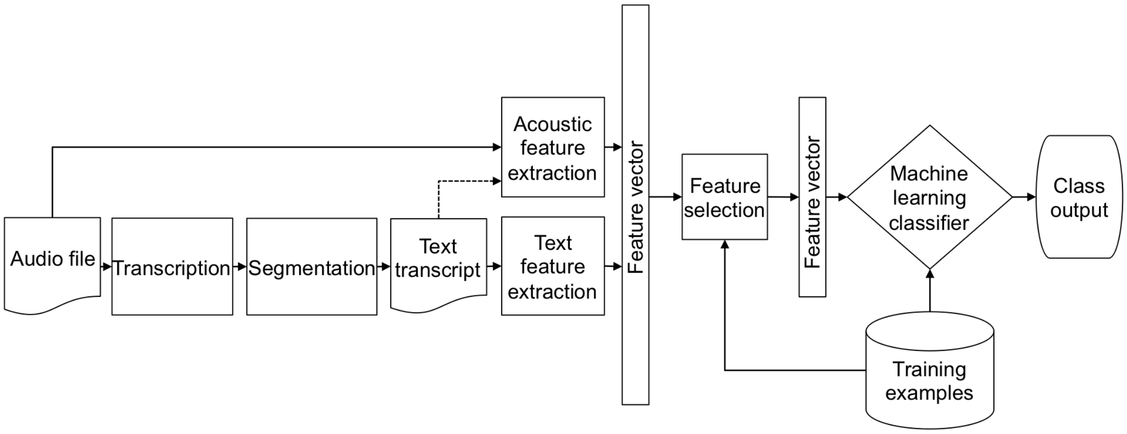


Figure 1.1: Processing pipeline for automatic detection of dementia from speech. The input to the system is an audio file containing a connected speech sample. Most studies use a human transcriber for the "Transcription" and "Segmentation" steps. Many studies omit the "Acoustic feature extraction" step. Note that the "Text feature extraction" step itself can contain multiple sub-processes, such as parsing, chunking, tagging, etc.

for the extracted features across groups, but do not explain in any detail what those values might imply about language changes due to cognitive decline. Garrard et al. (2014) are unable to identify a clear pattern in the features distinguishing between dementia pathologies, likely in part due to the sparse nature of the features (lexical unigram frequencies). Other studies employ automated techniques to extract relevant features, but then stop short of actually using those features to predict class membership (Pakhomov et al., 2010b,a; Garrard and Forsyth, 2010; Meteyard et al., 2014). Here, in contrast, my objective is to find clinically motivated features that both lead to high classification accuracies and reveal interpretable information about the dementia syndrome in question.

Furthermore, it is rarely addressed how the different stages of the NLP processing pipeline (Figure 1.1) affect the data and the resultant extracted features. Very few studies consider the practical questions of automatic speech recognition and segmentation at all, let alone examine the effect of these noisy processes on feature values, or investigate how these processes may disproportionately affect data from different patient populations. Peintner et al. (2008) use automatic speech recognition (ASR) to generate transcripts for dementia participants, listing word error rates ranging from 30% to 61% (and up to 100% for one participant). However, there is no discussion of how the lexical features and part-of-speech tags may have been affected. In related work, Jarrold et al. (2014) use the same speech recognition system, although they do not report word error rates. They do list several hypotheses based on the dementia literature, and they also compare their findings for these hypotheses using both manual and automatic transcriptions. However, although they report different results on the two sets of transcripts, there is no analysis of the conflicting results or discussion of whether some features are more affected than others. Lehr et al. (2012) report word error rates of 36–47%, and find that increased error rates led to decreasing feature accuracy, although they consider the somewhat different task of scoring story re-tellings for recall of relevant items, rather than assessing linguistic ability. Lehr et al. (2013) report an improved word error rate of 25.6% but similarly seek to assess story recall, rather than linguistic ability. None of these four studies attempts to measure the syntactic complexity of the narrative samples. To the best of my knowledge, all other studies of automated classification of dementia from narrative speech use manual transcripts.

In some machine learning applications, classification accuracy may be the most relevant outcome of a study. Accuracy is also important here. However, in healthcare applications, the features which go into a classifier may be just as important to the end user as the result that comes out. In this dissertation, I aim to show that *linguistic features automatically extracted from the connected speech samples of individuals with dementia can both differentiate these individuals from healthy, older controls and contribute to our knowledge of the nature of language impairment in dementia.* As a secondary goal, I aim to address the challenges and ultimate feasibility of a fully automated processing pipeline.

Background information about language impairment in dementia is given in Chapter 2. In this research, I focus on a dementia syndrome called primary progressive aphasia (PPA). Aphasia is an acquired language disorder, and in PPA the aphasic symptoms are due to an underlying neurodegenerative disease, principally affecting the language areas of the brain. Because people with PPA are relatively cognitively intact, PPA speech represents an ideal source of data on which to test linguistic analyses in the absence of other impairments. Furthermore, there are different variants of PPA that affect semantic and syntactic production differentially, allowing us to test whether certain features are more sensitive to semantic or syntactic abnormalities. I also provide information about related manual and computational studies of connected speech in PPA and other types of dementia. Additionally, I present some of the issues around measuring the syntactic complexity of speech, in which, unlike writing, dysfluencies are common and sentence or utterance boundaries are unclear. Finally, I expand on the potential applications of this work and discuss some of the ethical concerns related to automated detection and monitoring of dementia.

The objective of the work discussed in Chapter 3 is to automatically extract linguistic information from the PPA data, train machine learning classifiers to differentiate between the different variants of PPA relative to healthy controls, and interpret the selected features in the context of previous studies of connected speech in PPA. I start by considering only linguistic features derived from the speech transcripts. I then incorporate acoustic features derived from the associated audio files. The classifiers achieve reasonable accuracies in this preliminary work, but the features do not identify syntactic differences between the two subtypes. Therefore, in the second half of this chapter I examine a more fine-grained set of syntactic features and compare the different distributions of these features across groups. This chapter ends with an ablation study to determine the optimal set of features on which to train the classifiers, which ultimately achieve accuracies of over 90% in each task.

Having shown that the extracted features can reliably differentiate the groups, I turn to the question of how these features degrade in the face of the noisy processes which will necessarily be part of a fully automated system. In the first part of Chapter 4, I experiment with using off-the-shelf speech recognition software, and find that it does a poor job of recognizing these data. However, despite the high error rates, some features are still relevant, particularly those that measure the psycholinguistic properties of words, such as frequency and familiarity. By combining these relatively robust features with the acoustic features, which are unaffected by the speech recognition process, the classifier can still distinguish between PPA and control participants with over 80% accuracy, although distinguishing between the subtypes is more difficult. In the second part of the chapter, I consider the effect of automatic boundary segmentation on traditional measures of syntactic complexity. Although the accuracy of the segmentation is only moderate, relative to the manual segmentation, we actually observe higher classification accuracies using the automatically segmented transcripts. This illustrates the potential divergence between accurate classification and accurate feature values, although some examples from the transcripts also call into question the assumption that human-generated annotations are, by definition, more accurate.

While Chapter 4 addresses primarily practical concerns, Chapter 5 focuses on a question of more theoretical interest: Is there a difference in the speech patterns of individuals with agrammatism due to PPA versus agrammatism due to post-stroke aphasia? The two existing studies on this topic report conflicting results (Patterson et al., 2006; Thompson et al., 2013). In this case, the use of machine learning classification is not directed towards a clinical application (the onset of aphasia after a stroke is sudden and would not be confused for the slow, insidious onset of PPA), but rather to explore the question of whether a detectable difference between the groups exists. A best accuracy of 76% is achieved, suggesting that there are differences between these two populations. An analysis of important features reveals differences in the usage of verb tenses and prepositional phrases.

Having validated the methodology on PPA, in Chapter 6 I examine whether a similar analysis will detect Alzheimer's disease from speech samples. Alzheimer's disease (AD) is the most common form of dementia, and is marked by a gradually worsening memory impairment. Over time, other cognitive areas are also affected. While language impairment is not the primary symptom of AD, numerous studies have suggested that changes in language can occur very early in the disease, and even in the prodromal phase (Forbes et al., 2001; Ahmed et al., 2013; Cuetos et al., 2007). By including additional features to capture the expected lexical content of the narratives, classification accuracies of up to 81% are achieved. A factor analysis reveals the different underlying language dimensions that are affected. Finally, the experiment is repeated using ASR transcripts rather than manual transcripts, and we find that the resulting accuracy is within one standard deviation of the system's performance on manual transcripts, and that many of the relevant features are still significantly different between the groups. This suggests that a completely automated system may become feasible as speech recognition accuracy improves.

In Chapter 7 I summarize and discuss the results, and list some limitations of the work and how they might be improved. I also present a number of areas for future work, building on the foundation presented here.

Much of this research has already been published, and all of it was collaborative. The stereotype of the 'lone wolf' computer scientist does not translate to such intensely multidisciplinary work. Therefore, I have indicated with a footnote all sections containing previously published material, and the co-authors of the material. These publications include Fraser et al. (2013a,b, 2014a,c, 2015a,b). I have received permission from each of my co-authors to reproduce the work here, and all of the publishers permit the re-use of material by the author in an academic dissertation. The pronoun *we* is used throughout these publications, and for consistency throughout the unpublished sections as well, with the exception of the introductory and concluding chapters.

Chapter 2

Background and related work

There are many different types of dementia that can exhibit language symptoms, including Alzheimer's disease, frontotemporal dementia, dementia with Lewy bodies, and vascular dementia (Rohrer et al., 2008). In this dissertation, I will focus on the syndrome of primary progressive aphasia (PPA), although an extension to Alzheimer's disease is presented in Chapter 6. Individuals with PPA are specifically impaired with respect to language, while other cognitive abilities are generally spared. This makes PPA speech a natural choice of data set on which to test speech and language measures. More details about PPA and its subtypes are presented below.

In this chapter, I also discuss previous work using connected speech samples to detect and classify dementia. In the past, this type of analysis was carried out by hand, although in recent years there has been growing interest in using tools from natural language processing to automate existing processes and to develop new methods of analysis. I focus on measures of syntactic complexity and the challenges associated with applying these metrics to speech data, including the problem of utterance segmentation in speech. I then give a brief overview of automatic speech recognition, particularly as it has been applied to ageing voices, and finish with a discussion of some of the potential applications of this research as well as the ethical issues surrounding them.

2.1 Primary progressive aphasia

Primary progressive aphasia is a language disorder which occurs as the result of neurodegeneration. It progresses slowly and insidiously. According to the international consensus criteria, a diagnosis of PPA requires the presence of aphasic symptoms in the absence of other major cognitive deficits (such as memory or executive function) or behavioural disturbances (Gorno-Tempini et al., 2011).

PPA is often considered to be a type of frontotemporal dementia (FTD), along with a behavioural variant (bv-FTD) (Harciarek and Kertesz, 2011). The bv-FTD subtype involves a change in personality or behaviour, and is not normally associated with language symptoms, although it has been suggested that patients with bv-FTD may develop difficulties at higher levels of language functioning, such as producing coherent, organized discourse (Ash et al., 2006).

PPA may be broken down into subtypes based on specific language deficits. In early work, a distinction was drawn between "fluent" and "nonfluent" subtypes, roughly corresponding to the distinction between Wernicke's aphasia and Broca's aphasia in classical aphasiology (Leyton and Hodges, 2014). However, this view is now regarded as overly simplistic. Due to the range of different underlying causes for PPA, and debate regarding the diagnostic criteria for each subtype, subtype classification can be difficult in some cases. While there are broad patterns of decline, individual patients often show at least some symptoms from different subtypes (Rogalski et al., 2011). Nonetheless, most cases of PPA can be assigned to one of three subtypes: semantic variant (sv-PPA), nonfluent/agrammatic variant (nfv-PPA) and logopenic variant (lv-PPA) (Gorno-Tempini et al., 2011).[1]

The semantic variant, sometimes called "semantic dementia", is marked by fluent but empty speech, anomia, deficits in comprehension, and spared grammar and syntax (Mesulam et al., 2012). Patients with sv-PPA often have particular difficulty with nouns as opposed to verbs,

[1] Since no abbrevations for the subtypes are given in Gorno-Tempini et al. (2011), here we follow the abbreviation conventions of Leyton and Hodges (2014).

particularly low-frequency nouns (Rohrer et al., 2008). In general, semantic noun categories (e.g., animate versus inanimate objects) are affected more or less equally, although there have been a few cases where category effects were seen (Lambon Ralph et al., 2003a). In conversational speech, sv-PPA patients may avoid these low-frequency words by using circumlocutions or higher-frequency hypernyms (Rohrer et al., 2008). They may also use more pronouns or other closed-class words in place of more specific words (Harciarek and Kertesz, 2011).

The nonfluent/agrammatic variant is characterized by halting, agrammatic speech, reduced syntactic complexity, word-finding difficulties, and relatively spared single-word comprehension (Mesulam et al., 2012). The agrammatism is usually less severe than in Broca's aphasia (Harciarek and Kertesz, 2011; Wilson et al., 2010). This subtype is called "nonfluent" because of the typically slow, effortful speech of nfv-PPA patients. However, recent work by Thompson et al. (2012) shows that levels of fluency and grammatical abilities vary from individual to individual and can be dissociated. Patients may also show altered prosody and speech sound errors or apraxia of speech (Grossman, 2010; Harciarek and Kertesz, 2011). As the disease progresses, the patient may become mute (Wilson et al., 2010).

The third subtype, lv-PPA, was identified by Gorno-Tempini et al. (2004) after their experience with PPA patients who did not seem to fall into either of the above categories. The two main features of lv-PPA are word-finding difficulties and impaired repetition (i.e., difficulty repeating a sentence or phrase spoken by the examiner), as well as speech errors. Usually single-word comprehension is spared, along with motor speech and grammar (Gorno-Tempini et al., 2011), although speech is slow and syntactically simple (Gorno-Tempini et al., 2004).

Neuroimaging can be a useful diagnostic tool, although the mapping from clinical subtype to location of brain pathology is not always direct. Mesulam et al. (2009) observe that PPA in general is marked by a "distinctly asymmetric atrophy" centered in the left hemisphere language network. Harciarek and Kertesz (2011) note that while most neuroimaging of nfv-PPA patients shows atrophy around the left posterior inferior frontal gyrus, some patients show more generalized atrophy, and others show none. The sv-PPA subtype is associated with bilateral atrophy in the anterior temporal lobes, usually greater on the left, eventually spreading to the ventral and lateral temporal lobes (Harciarek and Kertesz, 2011). Neuroimaging of lv-PPA patients shows damage to the left posterior superior temporal lobe, as well as parts of the parietal lobe (Harciarek and Kertesz, 2011). Similar patterns of degeneration are seen in Alzheimer's disease, and there is some debate over whether lv-PPA should be considered a subtype of FTD or a variant of Alzheimer's disease (Munoz et al., 2007; Ahmed et al., 2012).

2.2 Detecting PPA and related disorders from connected speech

The diagnosis of language-specific syndromes is usually made on the basis on a series of neuropsychological tests, family history, interviews with the patient and their relatives, and the overall impression of the clinician (Pakhomov et al., 2010a). The tests will often include tests of language production such as naming, reading, and writing, as well as comprehension tests. Performance in each of these language areas is an important factor in making the diagnosis. However, Rohrer et al. (2008) write that "[s]ystematic analysis of an extended sample of the patient's spontaneous (propositional) speech is the single most valuable aspect of the examination" when a patient presents with word-finding difficulties. Billette et al. (2015) state that in the diagnosis of PPA, "[l]inguistic analysis of connected speech is the gold standard but is impractical outside the research setting." Additionally, in the early stages of PPA or Alzheimer's disease, traditional cognitive testing may not be sensitive enough to detect small changes that can be seen in normal conversation (Bucks et al., 2000). Careful analysis of connected speech can therefore provide valuable information regarding an individual's language capacities.

Prins and Bastiaanse (2004) discuss both the clinical and theoretical benefits of analyzing spontaneous and semi-spontaneous speech from aphasic patients, but they also present some caveats to the approach. In particular, they mention the difficulty of interpreting the results of such analysis, when performance can vary over different sessions or even within a single session, and given the complex relationships between the different linguistic levels that contribute to the production of narrative speech. Bearing these issues in mind, they stress the importance of speech analysis as a realistic and objective measure of language processing capability, in individual cases as well as in group studies.

In practice, narrative speech production is analyzed using one of two general procedures (Pakhomov et al., 2010a). A trained professional may observe the patient in conversation or on a more structured speech task, and rate their performance subjectively (e.g., on a scale from “unimpaired” to “severely impaired”). Or, the speech may be transcribed, and the patient’s performance characterized through the calculation of objective measures of linguistic ability. The former method is more common in a clinical setting, given the constraints of time and resources. The latter method is traditionally performed by hand in a research setting and, according to Pakhomov et al. (2010a), is “likely to produce more objective and reproducible results.”

2.2.1 Manual analysis

One popular method for performing in-depth analysis of aphasic speech is called Quantitative Production Analysis, or QPA (Saffran et al., 1989). In QPA, speech transcripts are first edited and segmented into utterances, then analyzed for a number of different syntactic and semantic features. This analysis has been used, for example, to identify the differences between fluent and non-fluent aphasic speech (Bird and Franklin, 1996) and to describe the qualities of narrative production in Broca’s aphasia (Rochon et al., 2000).

Wilson et al. (2010) used QPA to analyze narrative speech in PPA. They analyzed speech samples (elicited through a picture description task) from 50 participants with PPA, as well as participants with bv-FTD and healthy older controls. They reported that all of the PPA patients made syntactic errors, although these errors were more frequent in the nfv-PPA group. The nfv-PPA group also had the lowest speech rate and produced the most false starts and filled pauses. In the sv-PPA group, the proportion of closed-class words was increased, as well as the ratio of pronouns to nouns and verbs to nouns. Interestingly, participants in the sv-PPA group produced more embedded phrases than controls, as a result of their word-finding difficulties.

Another formalized method for narrative speech analysis is the Northwestern Narrative Language Analysis (NNLA) (Thompson et al., 1995). NNLA involves five levels of analysis, from transcription, to determining utterance-level grammatical correctness, to detailed analysis of verb complexity and verb argument structure. It is well-suited to studying grammatical impairments, such as those seen in Broca's aphasia (Thompson et al., 1997b) and nfv-PPA (Thompson et al., 1997a, 2012).

As one example, Thompson et al. (2012) studied the relationship between fluency and agrammatism in PPA using the NNLA method. They analyzed speech samples from 37 PPA patients and 13 healthy controls, elicited through a story-telling task. They measured variables such as speech rate, mean length of utterance, proportion of grammatically correct sentences, ratio of open- to closed-class words, ratio of nouns to verbs, and correct production of verb inflections and noun morphology. They found that some PPA patients showed reduced fluency without agrammatism, primarily in the logopenic variant, although also in some cases of sv-PPA.

Other researchers have focused on different aspects of narrative speech. Ash et al. (2006) measured global and local connectedness in a picture-based story-telling task, and found that bv-FTD patients were more impaired on these higher-level discourse measures than sv-PPA patients (who struggled more with finding the correct words) and nfv-PPA patients (who produced the shortest narratives). In later work, Ash et al. (2010) analyzed the specific nature of speech errors in narrative speech from nfv-PPA participants, drawing a distinction between *phonemic* errors (resulting in permissible sequences of phonemes in English) and *phonetic* errors (resulting in phoneme sequences not permissible in English). They found that most speech errors were phonemic, suggesting the primary impairment in nfv-PPA is not a motor-planning deficit.

Considering nfv-PPA specifically, Graham et al. (2004) used a picture description task to elicit speech from participants with nfv-PPA. They reported reductions in narrative length and speech rate, as well as information content, relative to controls. Most of the participants did not show signs of agrammatism. Knibb et al. (2009) analyzed conversational speech from participants with nfv-PPA, finding increased speech sound errors and grammatical errors, and simplified (but not telegraphic) syntax.

For sv-PPA, Meteyard and Patterson (2009) analyzed errors in autobiographical interviews from 8 participants with sv-PPA and 8 matched controls. They found that sv-PPA participants did make more errors than controls; specifically, they would omit or substitute open-class items (e.g., nouns or verbs), and substitute but not omit closed-class items (e.g., function words). There was little evidence for frank agrammatism or phonological errors. Sajjadi et al. (2012) compared two types of connected speech, a picture description task and a semi-structured interview, from participants with sv-PPA and Alzheimer's disease. They found that the impairment in sv-PPA was less obvious in the interview task, as participants had more control over the topic of conversation and could avoid difficult words. Semantic impairments were more apparent in the picture description, but the circumlocutions were still rare. In contrast, abnormalities in syntax and morphological processing occurred more frequently in the structured interview.

To summarize, there are a number of previous studies reporting manual analysis of connected speech in PPA. Some of these studies used a formalized system of analysis, such as QPA or NNLA, while others developed project-specific methodologies. The speech elicitation task varied, but typically involved either a story-telling task (from memory or aided by a series of pictures), a description of a single picture, or a directed interview. Variation in the results may be partly due to the manner in which the speech was elicited (Sajjadi et al., 2012). In general, participants with sv-PPA produced an increased proportion of closed-class words, pronouns, and verbs, showed evidence of word-finding difficulty, and did not exhibit notable difficulties with grammar or motor speech. Participants with nfv-PPA typically had a slower speech rate and produced shorter narratives with more false starts, filled pauses, and phonemic errors. The frequency of agrammatism in nfv-PPA varied across studies.

2.2.2 Computerized analysis

The manual systems of analysis described above have formed the basis of our understanding of narrative speech in PPA. However, due partly to the time-consuming nature of the analysis, there has been interest in automatic methods for feature extraction and classification of aphasic speech. Beyond the need for efficient information extraction, the application of machine learning allows for a new type of analysis: determining the most likely diagnosis on the basis of linguistic features.

In the case of PPA specifically, Pakhomov et al. (2010b) analyzed manual transcriptions of connected speech samples from participants with FTD. They calculated the perplexity and out-of-vocabulary rate for each transcript, and compared the differences across patient groups. Both measures were sensitive to the impairments seen in sv-PPA. In a related study, Pakhomov et al. (2010a) extracted both linguistic and acoustic features, and found a statistically significant difference between some FTD subtypes on the basis of pause-to-word ratio, frequency of dysfluent events, and pronoun-to-noun ratio. Neither of these studies attempted to classify patient narratives on the basis of the extracted features.

Meteyard et al. (2014) used two available software packages to analyze autobiographical interviews from 8 patients with sv-PPA and 8 healthy controls. They compared the distributions of nouns and verbs across the two groups, as well as the distributions of a variety of different syntactic structures. They found that participants with sv-PPA used a reduced range of complex morpho-syntactic forms. They also reported good agreement between the automatic analysis and the hand-coded analysis, although they attributed some issues to the software's inability to distinguish between auxiliary and regular verbs.

Garrard and Forsyth (2010) analyzed picture description narratives from 21 participants with sv-PPA and 21 control participants by counting the frequency of occurrence of each lexical item, and then conducting a principal components analysis (PCA). When the first two components were plotted, the controls formed a relatively tight cluster, while the sv-PPA cluster was more diffuse, and partly overlapped the control cluster. The authors interpreted the first component as corresponding to discourse structure (positive loadings on grammatical function words and negative loadings on deictic words and off-topic utterances), and the second component as corresponding to differences in semantic content (positive loadings on pronouns, negative loadings on determiners and content-bearing nouns and verbs). Again, no automated method of assigning unseen patient narratives to one cluster or the other was tested. However, in a subsequent study, Garrard et al. (2014) used the unigram frequency information to automatically classify sv-PPA versus controls, as well as to classify narratives from sv-PPA participants with left- versus right-temporal lobe predominant atrophy. In the former case, the important features included low-frequency content words, generic words, and extra-narrative utterances such as “you know”. In the latter case, only a very few words were selected as significant, reflecting the similarity between the two groups.

One study which is particularly relevant here is that of Peintner et al. (2008), in which the authors used automatic speech recognition software to create transcripts of connected speech samples from participants with FTD, and then classified the samples using lexical and acoustic features. A subsequent paper from the same group used manually transcribed samples to detect signs of cognitive impairment, depression, and Alzheimer’s disease (Jarrold et al., 2010). These papers report relatively good classification accuracy, but contain very little discussion of the features which were selected and how they might relate to the disorders being studied. Jarrold et al. (2014) present a somewhat more nuanced approach, in which they use the automated analysis pipeline to explore certain language hypotheses based on the clinical literature. In general, they find that their automated approach produces results consistent with previous, manual analysis.

Other related work has applied natural language processing techniques to different degenerative disorders. Roark et al. (2011) used similar techniques to measure the speech characteristics of participants with mild cognitive impairment (MCI), a clinical syndrome that can precede Alzheimer’s disease. The elicitation task that they used involved the speaker retelling a story that had been told by the examiner, so there was a significant memory component to the task as well. They found that by using a combination of linguistic test scores and automatically calculated speech and language features, they were able to achieve a better classification result than by just using test scores alone.

Other researchers have explored using only information from the speech signal, without the transcript, to detect signs of cognitive impairment. D'Arcy et al. (2008) extracted a number of syntactic-temporal features, such as the number and duration of pauses, from read and spontaneous speech samples. Using the features extracted from read speech, they were able to predict with an accuracy of 76.7% whether the speaker had scored below a threshold on a test of cognitive function. In a subsequent paper from the same group, it was shown that similar features could be extracted over telephone recordings, suggesting that this type of analysis could be performed remotely (Rapcan et al., 2009). Tóth et al. (2015) also used acoustic features (including articulation rate, speech rate, utterance length, pause duration, number of pauses, and hesitation rate) to distinguish between 32 participants with MCI and 19 elderly controls with a best accuracy of 80.4%.

A slightly different approach was taken by Hakkani-Tür et al. (2010), who used ASR and natural language processing methods to predict numerical cognitive test scores rather than participant categories. They achieved a high correlation with manual scores on a picture description task and story retelling task by automatically identifying and counting semantic units. Their participants included a mix of younger and older people, but no participants with a clinical diagnosis of dementia.

There has also been growing interest in detecting signs of Alzheimer's disease from speech. This literature is discussed in Chapter 6. Finally, there has also been work presenting computational analyses of speech and language in other cognitive and neuropsychological impairments, including (but not limited to) Parkinson's disease (Tsanas et al., 2012), amyotrophic lateral sclerosis (Yunusova et al., 2016), drug-induced cognitive impairment (Pakhomov et al., 2013), autism (Prud'hommeaux et al., 2011, 2014), and specific language impairment (Gabani et al., 2011; Solorio, 2013). Studies of this type are generally beyond the scope of this thesis.

2.3 Machine learning for text classification

As mentioned in the previous section, a crucial distinction between much of the manual analysis and the more recent computational work is that the focus has shifted from merely describing the samples to also automatically labelling them with the most probable diagnosis. If we consider only the speech transcripts as input, then the problem of determining the diagnosis given the text can be thought of as a text classification problem. Text classification is a very broad area of research in natural language processing. Other similar problems in this category include: authorship attribution (Juola, 2006), authorship profiling (Argamon et al., 2009), automatic essay grading (Dikli, 2006), determining the reading level of a text (Petersen and Ostendorf, 2009), and scoring language-learners' speech fluency (Zechner et al., 2009). What these diverse applications have in common is the need to identify and measure vocabulary richness, lexical choices, syntactic complexity, the amount of information conveyed, clarity, coherence, grammaticality, and many other patterns of language use. Many of the features used in such applications have also proven useful in the task of dementia detection, as we will see throughout.

The core goal of these applications is to determine the most likely class to which a text belongs (e.g., the author of an unknown play, or the grade of an unmarked essay). This is typically accomplished through supervised machine learning. In this thesis, I consider a variety of different classification algorithms, which are summarized below.

Naïve Bayes

Naïve Bayes is a classifier based on Bayes's theorem. It is called "naïve" because it makes the strong simplifying assumption that all of the features are conditionally independent given the class. The classifier learns estimates for the class-conditional probabilities and priors for each class from the training data. In the classification stage, it uses Bayes's theorem to assign each data point to the class that maximizes the posterior probability. Naïve Bayes is widely used, even in cases where the independence assumption is known to be false, and often performs well. The rationale for this is that even though the probability estimates may be inaccurate, the classification results (which depend only on which probability is the highest, and not on the actual numbers) can still be good (Manning et al., 2008).

Logistic regression

In contrast to naïve Bayes, which attempts to model the classes themselves, logistic regression is a discriminative classifier which attempts to model the boundary between the classes instead. Logistic regression estimates the posterior probability directly from the training data. Research suggests that naïve Bayes may perform better in cases where there is not a large amount of training data (Ng and Jordan, 2002). However, the benefit of logistic regression is that it does not assume the features are conditionally independent. Peintner et al. (2008) used logistic regression, along with two other classifiers not considered here, on various classification tasks involving FTD subtypes and healthy controls. They had mixed results, with logistic regression achieving the best results in two out of six cases.

Support vector machines

Support vector machines (SVMs) are another type of linear discriminative classifier which have become very popular in natural language processing applications in the past several years (Manning et al., 2008). SVMs are maximum margin classifiers, which means they find the decision boundary between two classes that maximizes the margin between the two classes. In other words, they maximize the distance between the decision boundary and the nearest data points. If the data are not linearly separable, then the algorithm tries to maximize the margin while also minimizing the misclassification error (Manning et al., 2008).

Random forests

Random forests are a type of ensemble classifier consisting of several decision trees (Breiman, 2001). Each decision tree outputs a class prediction, and the output of the entire forest is the class that receives the most votes. Each individual tree is trained on a random subset of the training data, and a random subset of the features. This process is repeated for each tree in the ensemble (both training examples and features can be reused in different trees). Some benefits to random forests are that they can handle missing data gracefully, and easily extend to multi-class classification (Pal, 2005).

The algorithms discussed here are all standard machine learning algorithms in text classification. Recently, the concept of "deep learning" has become very popular, and deep learning algorithms have been successfully applied in a number of applications. However, these models are complex, and typically require large quantities of data in order to be trained effectively (although the problem of overfitting can be reduced to some extent by employing a regularization method, such as dropout (Srivastava et al., 2014)). In the work reported here, we have a very small amount of data from a machine learning standpoint, and so we restrict our analysis to relatively simple classification algorithms.

2.4 Syntactic complexity of speech

One aspect of language that we would like to measure automatically is the appropriateness and complexity of syntactic structures. *Agrammatism* refers to the inability of some aphasic speakers to communicate in a grammatically complete and correct fashion. It usually involves the omission or substitution of function words and morphological markers for tense, person, number, and gender (Thompson and Bastiaanse, 2012). One form of agrammatism is known as "telegraphic speech". In telegraphic speech, the form of the sentence remains relatively intact, but with most of the inflections and function words omitted, such that content words such as nouns and verbs are simply strung together with no supporting structure (Goodglass and Kaplan, 1983).

Although agrammatism is one of the core symptoms of nfv-PPA (Gorno-Tempini et al., 2011), there is debate over the true extent of its occurrence in PPA (Graham et al., 2004; Knibb et al., 2009; Wilson et al., 2010; Thompson et al., 2012). Thus, exploring different measures of syntactic production and comparing them across diagnostic groups will be a key aspect of this work.

In less severe cases, sentences may be grammatically correct, but with greatly simplified syntax. Simplified syntax (including the production of short, simple sentences, a reliance on canonical sentence structures, and a reduction in embedded clauses) is often associated with agrammatism, but the two can be dissociated (Nadeau, 2012). Simplified syntax has been observed in all subtypes of PPA (Wilson et al., 2010). However, reduced syntactic complexity is not necessarily the result of pathological language impairment; Cheung and Kemper (1992) compared the syntactic complexity of spoken narratives from older and younger adults and found that sentence complexity does tend to decline with age, particularly in terms of sentence length and number of embeddings, and suggested that the effect is related to normal age-related decline in working memory. Similarly, although there is evidence of syntactic decline in Alzheimer's patients, it has been argued that such effects are primarily due to memory deficits rather than true grammatical impairments (Reilly et al., 2011).

One simple way to measure syntactic information is to count the part-of-speech (POS) of each word in the sample. For example, QPA calculates the proportion of closed-class words, the ratio of determiners to nouns, the ratio of pronouns to nouns+pronouns, and the ratio of verbs to nouns+verbs (Saffran et al., 1989; Rochon et al., 2000). These measures can be good discriminators between fluent and nonfluent aphasia types, since fluent patients (e.g., sv-PPA) tend to produce more pronouns and verbs, and nonfluent patients (e.g., nfv-PPA) tend to omit determiners and produce more nouns. POS tagging also makes it possible to count the number of adjectives, which are less likely to occur in nonfluent speech (Nadeau, 2012). These quantities are easily computed automatically with an automatic POS-tagger, such as the Stanford tagger (Toutanova et al., 2003). QPA also measures such quantities as the proportion of nouns which require a determiner and actually have one, and the number of inflectable words that are actually inflected (Saffran et al., 1989). These measures are more difficult to compute automatically, since they require a grammaticality judgement rather than simply counting the frequency of production.

Other syntactic measures assess complexity beyond the single-word level. Cheung and Kemper (1992) examined 11 complexity measures in their study of older adults' speech. They considered mean length of utterance (MLU), mean number of clauses per utterance (MCU), Developmental Sentence Scoring (DSS; a method to assess grammatical development in children based on their use of eight different grammatical structures), Index of Productive Syntax (IPSyn; a scale for grammatical development in children based on 56 target grammatical types), Developmental Level (DLevel; a method to assess grammatical competence based on seven (or eight) target sentence constructions), Directional Complexity (DComplexity; a measure of linguistic difficulty of a text based on the presence of various patterns and structures), Propositional Density (PDensity; a measure of the amount of semantic content in a text), two variants of Yngve depth (a tree-based measure that measures how left-branching a parse tree is), and two variants of Frazier count (a tree-based measure in which each node in a parse tree is given points based on the length of the path either to the root or to the lowest node with a left sibling). They found significant differences between older and younger adults on all measures except MLU, IPSyn, and PDensity (Cheung and Kemper, 1992).

Some of these complexity metrics have been automated and used in computational analyses of impaired speech and language. Meteyard et al. (2014) automatically calculated Dlevel for speech samples from sv-PPA patients and controls, using a software package called ShaC (Voss, 2005). However, although they found differences in the fine-grained syntactic measures that were computed as input to the DLevel analysis, there was no difference between the groups on DLevel itself.

Roark et al. (2011) considered MLU, Frazier count, Yngve depth, and PDensity in their analysis of story retelling in mild cognitive impairment (MCI). They also considered the mean length of clause (MLC), content density (the ratio of open-class words to closed-class words), dependency distance (the distance between words connected by a dependency relation), and POS tag cross-entropy (a measure of the probability of a sequence of POS tags). They found that the difference between MCI and control groups on MLC and content density was significant in both the immediate and delayed retellings, POS cross-entropy was significant only in the immediate retellings, and Yngve depth and dependency distance were significant only in the delayed retellings.

One challenge in applying these methods to speech is that speech, unlike writing, is rarely produced in well-defined sentences, and the location of sentence boundaries in speech transcripts can be unclear. While written sentences are marked by punctuation and capitalization, these cues do not exist in speech. Instead we must rely on prosodic and lexical cues, although these are far from unambiguous. Coordinating conjunctions in speech can have a "loose discourse linking function" that does not indicate grammatical connection (Leech, 2000). This means that speakers can create extremely long utterances consisting of independent clauses connected by coordinate conjunctions, and it is not clear where one sentence ends and the next begins. Alternatively, an utterance can also consist of multiple clauses joined without the use of conjunctions, but uninterrupted by pauses or other prosodic cues that might indicate a sentence boundary. As Miller and Weinert (1998) write, "it is far from evident that the language system of spoken English has sentences, for the simple reason that text-sentences are hard to locate in spoken texts."

Miller and Weinert argue instead in favour of the clause as the primary unit of speech, and suggest that clause boundaries can be detected by locating a verb and its complements. Another popular position was put forth by Hunt (1966), who advocated the use of T-units as the basic unit of speech, where a T-unit contains an independent clause plus any attached dependent clauses. Hunt argues that there are two ways to add complexity: either by moving information from a coordinate clause to a subordinate clause (*The woman is tall and the woman went to the store* becomes *The woman, who is tall, went to the store*), or by moving information to the main clause (*The tall woman went to the store*). He claims that both of these types of complexity are captured by T-unit analysis, rather than by sentence analysis. Foster et al. (2000) extend the definition of T-unit in their *analysis of speech unit*, or *AS-unit*, by allowing AS-units to consist of a main clause or "sub-clausal unit" as well as any attached subordinate clauses. This modified definition is particularly relevant to conversational speech, which often contains fragments and elliptical expressions; Leech (2000) estimated that 30% of utterances in conversational speech are non-clausal.

Another potential issue which arises in speech is the presence of dysfluencies, such as filled pauses, repetitions, and false starts. The traditional approach has been to remove dysfluencies and other non-narrative speech before syntactic analysis (Saffran et al., 1989; Thompson et al., 1995). Dysfluencies can artificially inflate some syntactic complexity measures, such as mean sentence length or Yngve score. However, Garrard et al. (2014) found that what they called *paralinguistic* elements conveyed useful diagnostic information, particularly discourse markers (e.g., *you know*) and comments about the task itself (e.g., *I can't remember*).

In the following work, we are constrained in our choice of primary syntactic unit by the transcription protocols that were used when the data was transcribed, and the definition of a "sentence" or "utterance" varies somewhat across data sets. However, wherever possible, we consider syntactic analysis at multiple levels (e.g., by calculating mean length of clause, T-unit, *and* sentence). Regarding dysfluencies, we count filled pauses and false starts, then remove them for syntactic processing. However, with the exception of Chapter 5, we include all other transcribed speech in the analysis.

2.5 Automatic segmentation of speech

Once it has been decided which unit will form the basis of the syntactic analysis, the unit boundaries must be marked in the transcript. In manual analysis, this is typically done by the transcriber during the process of transcription; however, as Elvevåg et al. (2010) note, "the use of measures representing structure imposed by the transcriber ... is certainly not ideal." That is, a transcriber might affect the results of the analysis by introducing his or her own bias about what counts as a sentence or clause onto the data. To avoid this, transcribers are often given handbooks with explicit instructions on how to segment the text. Even so, there will be ambiguous cases and inter-rater agreement can vary. Reed et al. (2001) reported inter-rater reliability for utterance segmentation ranging from 73.3% to 99%, depending on the definition of "utterance". They also reported significant differences on their measures of syntactic complexity, depending on which definition for utterance was used.

Aside from the issue of agreement, it is clear that for fully automated analysis we will require a computational method for detecting syntactic units. Most previous work has focused on sentence-boundary detection. In many cases, the problem is framed as a binary labelling problem, where the boundary between two consecutive words must be labelled as either a sentence boundary or not a sentence boundary (Ostendorf et al., 2008).

One approach is to insert sentence boundaries according to prosodic cues, such as pauses. However, as discussed above, pauses are neither necessary nor sufficient evidence for sentence boundaries. Particularly in the case of PPA, we might expect there to be a number of pauses which occur when the speaker experiences word-finding difficulties (also known as *hesitation pauses*), rather than just at syntactic boundaries (or *juncture pauses*). Other prosodic features which may help identify sentence boundaries include phone duration and pitch and energy features. For example, the last syllables before a sentence boundary may be lengthened, or the pitch may slowly decrease over the course of a sentence (Kolář, 2008).

Sentence boundaries may also be labelled on the basis of lexical cues only. This can be done by searching for "trigger" words, such as conjunctions, that might indicate the start or end of a sentence (Gavalda et al., 1997), or using word or POS *n*-gram modelling (Stolcke and Shriberg, 1996; Stevenson and Gaizauskas, 2000). However, it is not surprising that the most successful algorithms combine both lexical and acoustic information (Shriberg, 2005; Ostendorf et al., 2008).

There has not been as much work on the automatic detection of units other than sentences. One intuitive approach would be to detect the sentence boundaries, then build parse trees of the sentences, and search the resulting trees for clauses or T-units. Indeed, Stevenson and Gaizauskas (2000) wrote, “it is difficult to imagine how clauses could be identified without parsing.” Of course, this method assumes that the sentence boundary detector has done a good job, and there are no clauses which cross the proposed sentence boundaries.

Jørgensen (2007) proposed a novel approach to clause detection, in which he classified each coordinating conjunction as belonging to the syntactic level or the discourse level (building on the idea that conjunctions in spoken language are often used to perform discourse or pragmatic functions). However, his 2007 paper presents results only for the classification of coordinating conjunctions, and not the proportion of clause boundaries which are correctly labelled.

Another method is to build the clauses from the bottom up, by combining words into phrases and then combining phrases into clauses. For example, Abney (1990) described his Cascaded Analysis of Syntactic Structure (CASS) parser, one component of which was a “clause filter.” The clause filter identified clauses from patterns of noun phrases, verb phrases, and function words.

In Chapter 4 we test standard algorithms for utterance segmentation on impaired speech, using the manually annotated utterance boundaries as a gold standard. Such methods could be applied to data that had been manually-transcribed at the word level, to remove inter-rater variability in utterance segmentation. However, the more likely scenario is for these methods to be used to automatically segment the text output from an automatic speech recognition system.

2.6 Automatic speech recognition for the elderly and the cognitively impaired

Automatic speech recognition (ASR) software converts speech (audio signals) into text. If we consider the noisy channel model of speech recognition, in which a source word sequence X is observed as a noisy acoustic sequence Y, then the task of speech recognition is to produce the most likely word sequence X^* according to:

$$X^* = \arg\max_X P(Y|X)P(X) \tag{2.1}$$

The acoustic model describes $P(Y|X)$, or the probability of an acoustic sequence Y given a word sequence X. The language model describes $P(X)$, the probability of a word sequence X. The performance of an ASR system depends strongly on how well the acoustic and language models actually model the properties of the incoming signal. In elderly people with dementia, both the acoustic properties of the voice *and* the words that they use may not conform to typical patterns, resulting in poor ASR performance.

In general, the accuracy of ASR systems on elderly voices tends to decrease with the age of the speaker (Vipperla et al., 2008). Elderly voices typically have increased breathiness, jitter, shimmer, and a decreased rate of speech (Vipperla et al., 2008). Older speakers may also exhibit articulation difficulties, changes in fundamental frequency, and decreased voice intensity (Young and Mihailidis, 2010). The underlying physical changes may be related to changes in the size and periodicity of the glottal pulse, and changes to the internal control loops of the articulatory system (Wilpon and Jacobsen, 1996). These factors can result in speech that is less intelligible to both human listeners and ASR systems. For example, Hakkani-Tür et al. (2010) found that in automatic scoring of a speech-based cognitive test, their ASR system had a higher word error rate (WER) for healthy speakers over the age of 70 than for those under the age of 70, with WERs between 26.3% and 34.1% for the elderly speakers, depending on the task and the gender of the speaker, while the error rates ranged between 21.1% and 28.2% for the younger speakers.

Aman et al. (2013) not only reported worse WERs for elderly speakers (age > 65) than younger speakers, but found a much wider variance in WER as well. That is, some elderly speakers were much more difficult to recognize than others. When they considered possible explanations for this variance, the WER was found to correlate with a measure of dependence, or loss of autonomy, due to physical or cognitive decline. Speakers who had a higher level of dependence were more difficult to recognize. While none of the speakers were described as having dementia, it is certainly possible that this result would generalize to people with dementia.

Effective speech recognition can be further challenged by the presence of linguistic impairments such as those occurring in PPA. For example, Goldwater et al. (2010) described a number of different characteristics of words which tend to be mis-recognized, and found that words directly preceding dysfluent events are difficult to correctly recognize. Paraphasias and other "out-of-vocabulary" words will also cause problems for the language models in ASR systems. To our knowledge, there has only been one previous study reporting the results of automatic speech recognition with PPA speakers. Peintner et al. (2008) analyzed speech from patients with nfv-PPA and sv-PPA as well as bv-FTD. They achieved a WER of 37% for sv-PPA and 61% for nfv-PPA. They also tested a control group, which had an average WER of 20%. In related work, Lehr et al. (2012) reported WERs of 44.3–50.6% on participants with mild cognitive impairment, and 36.1–45.0% for age-matched controls. In subsequent work, they were able to reduce the mean WER to 25.6% using a combination of acoustic and language model adaptation (Lehr et al., 2013).

There has also been previous research on adapting ASR for populations with other impairments, such as motor speech disorders. For example, Mengistu and Rudzicz (2011) achieved a reduction in word error rate of 37% for dysarthric speakers by recognizing error patterns within individual speakers. However, it is not clear how well these findings can be applied to speech from individuals with dementia, whose difficulties stem primarily from a cognitive, rather than motor, impairment.

2.7 Potential applications and ethical considerations

While it seems unlikely for automated cognitive assessment to *replace* a human clinician's diagnostic process, more reasonable applications of this technology include (a) an automated screening tool to help flag potentially cognitively-impaired individuals for a more rigorous assessment, (b) a computerized neuropsychological assessment which forms one part of a clinician's diagnostic framework, and (c) a method of analyzing language abilities quantitatively over time (to measure response to medication, for example).

2.7.1 Cognitive screening

Screening tests are used to detect signs of illness in apparently healthy individuals. They are applied across an entire population that is deemed to be at-risk for the disease. An example of this is the use of mammograms to screen for breast cancer in women over the age of 50. Screening allows for early detection of diseases, but there is a cost-benefit analysis that must be performed before implementing a screening program for a given disease and population. The World Health Organization published a set of guidelines for screening that include the following ten criteria (quoted from Wilson and Jungner (1968)):

1. The condition sought should be an important health problem.

2. There should be an accepted treatment for patients with the recognized disease.

3. Facilities for diagnosis and treatment should be available.

4. There should be a recognizable latent or early symptomatic stage.

5. There should be a suitable test or examination.

6. The test should be acceptable to the population.

7. The natural history of the condition, including development from latent to declared disease, should be adequately understood.

8. There should be an agreed policy on whom to treat as patients.
9. The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole.
10. Case-finding should be a continuing process and not a "once and for all" project.

Much of the subjectivity in these criteria lies in the interpretation of the adjectives: when is a treatment considered *accepted*? What does it mean for a test to be *suitable*? While Wilson and Jungner provide some practical guidelines for defining these terms, the answers are not clear-cut. Accordingly, there is ongoing debate regarding the acceptability of population-wide screening for dementia in older adults, with most of the attention being focused on dementia due to Alzheimer's disease. Many arguments for and against population screening have been proposed (Boustani et al., 2003; Solomon and Murphy, 2005; Ashford et al., 2006, 2007; Le Couteur et al., 2013; Schicktanz et al., 2014; Calzà et al., 2015). These arguments are summarized in Table 2.1. In short, the proponents argue that if screening leads to increased rates of early diagnosis, patients will be able to access treatment at a stage when it is most effective, their participation in risky behaviours (such as driving a car) can be reduced, and their family and caregivers will generally be more prepared and experience less stress. The opponents of screening argue that screening does not necessarily lead to diagnosis, due to challenges in following up after a positive screen and the costs associated with testing, and that early diagnosis does not necessarily lead to better outcomes and can have a negative emotional impact.

Interpretation of the arguments is complicated by the fact that different studies report different results with respect to prevalence, rates of under-diagnosis, and effectiveness of treatments and social programs. Nonetheless, as new and better treatments become available, it is reasonable to assume that screening will become more commonplace, at least amongst high-risk populations. Some potential benefits of a language-based screening test include: the sensitivity of language to early cognitive impairment (Cuetos et al., 2007; Ahmed et al., 2013), the simplicity and naturalistic nature of narrative speech (Tomoeda et al., 1996), the ability to conduct

Reasons for screening
Some causes of cognitive impairment are reversible when detected early (Calzà et al., 2015)
Fewer than 50% of AD cases diagnosed (Solomon and Murphy, 2005)
Cholinesterase inhibitors for treatment of symptoms are most effective when started in early stage (Solomon and Murphy, 2005)
Early diagnosis reduces risk of adverse events (e.g., car accidents, falls, forgetting to take medications) (Ashford et al., 2007; Calzà et al., 2015)
Allows for psychological and social intervention while individual is still competent (Ashford et al., 2007)
Improved family understanding of patient's behaviour reduces anxiety/stress (Ashford et al., 2007)
Caregivers of treated patients have better outcomes than caregivers of untreated patients (Ashford et al., 2007)

Reasons against screening
Not enough evidence that early diagnosis leads to better outcomes (Boustani et al., 2003; Schicktanz et al., 2014)
Challenges to providing adequate follow-up after screening (Solomon and Murphy, 2005)
Variability in professional training/ability to perform screening (Solomon and Murphy, 2005)
Issues with getting informed consent (Solomon and Murphy, 2005; Schicktanz et al., 2014)
Screening tools have not been validated in full range of educational and socioeconomic levels (Boustani et al., 2003)
Early diagnosis of AD can lead to adverse social consequences (stigma) and psychological anguish (Le Couteur et al., 2013; Schicktanz et al., 2014)
Some evidence that general public does not understand or support screening (Martin et al., 2015)
Financial cost of full neuropsychological assessment for all positive screens (Le Couteur et al., 2013)

Table 2.1: Reasons for and against routine screening of elderly people for signs of dementia. While there is an important distinction between screening and early diagnosis (namely, that a positive screen can *lead to* early diagnosis but is not a diagnostic result in and of itself (Ashford et al., 2007)), most arguments assume that screening programs will lead to increased early diagnosis.

the assessment remotely, e.g., over the phone (Rapcan et al., 2009), and the ability to repeat the assessment with different stimuli to avoid a learning effect (Forbes-McKay et al., 2013).

These benefits assume that the screening will take place under the supervision of a qualified healthcare professional. In recent years, a growing number of websites have started offering online cognitive screening tests (Robillard et al., 2015). While there can be advantages to online screening, including increasing access to information, allowing individuals to monitor their own cognition, and providing motivation to seek professional help when warranted, there are also a number of medical and ethical concerns associated with these tests. These concerns include poor accuracy, lack of scientific validation, ethical lapses such as non-disclosure of conflicts of interest, and a general lack of support and information accompanying the delivery of potentially upsetting news (Robillard et al., 2015). In particular, there is evidence that commercial entities selling unregulated and alternative "treatments" for AD may use online screens to attract customers (Robillard, 2016). Other relevant work has shown that asymptomatic individuals who discover they are at elevated risk of AD via direct-to-consumer genetic testing can be profoundly negatively impacted by the information, with little perceived benefit (Messner, 2011).

2.7.2 Computerized cognitive testing

Another potential application of automated speech analysis is integration within a computerized neuropsychological assessment device (CNAD). Computerized versions of neuropsychological tests have been proposed as one solution to the problem of limited clinician time in the face of increasing demand for dementia testing. CNADs have been used in the U.S. military and for evaluation of sports-related head injuries since the 1980s (Parsey and Schmitter-Edgecombe, 2013), and there is growing demand for their use in the field of age-related cognitive decline (Gates and Kochan, 2015). However, there are a number of basic criteria which any assessment tool must meet before it can be used in clinical decision-making.

Bauer et al. (2012) published a joint position paper summarizing the views of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology on the topic of CNADs. While positive towards the potential benefits of computerized assessment, including the ability to quickly test a large number of patients, reduced assessment times, reduced costs for administration and scoring, automated data storage, and increased accessibility for individuals in remote areas, the authors also identified eight key issues surrounding such devices. These eight issues are summarized below:

1. **Device marketing and performance claims:** Information about what a device does, how it does it, and its safety and effectiveness must be clearly provided.

2. **End-user issues:** End-users (e.g., patient or clinician) and their competencies and skills must be clearly defined.

3. **Technical issues:** Locally installed CNADs must operate in a manner comparable to the versions of the device on which normative data were collected.

4. **Privacy, data security, identity verification, and testing environment:** Patient data must be protected to an appropriate degree.

5. **Psychometric development issues:** Information about reliability, validity, and clinical utility (e.g., accuracy, sensitivity, specificity) must be provided.

6. **Examinee issues: Cultural, experiential, and disability factors:** Normative information must be provided with regards to the appropriateness of the test for patients from different racial, ethnic, and educational backgrounds, patients of different ages, and patients with cognitive, motor, or sensory disabilities.

7. **Use of computerized testing and reporting systems:** An automated output report from a CNAD may supplement but should not replace a clinician's evaluation of a patient.

8. **Checks on validity of responses and results:** CNADs should detect and identify examinee non-compliance or lack of motivation (e.g., through internal measures of effort or by recommending additional, validated tests of effort to be administered concurrently.).

In a follow-up article three years later, Gates and Kochan (2015) stated that, with few exceptions, there had not been acceptable progress on these issues by the developers of CNADs. Thus it is clear that before the research described in this thesis could be put to use in clinical practice, a substantial amount of work would be required. However, the inclusion of natural language analysis in a CNAD could be highly valuable, as previous studies have remarked on the conspicuous absence of language testing in existing computerized batteries (Dede et al., 2015; Tierney and Lermer, 2009).

2.7.3 Longitudinal assessment

A third potential application of this research is as a tool to provide quantitative ratings of language ability over time, either to detect signs of incipient cognitive decline, or potentially to track the progress of a therapy program or the effectiveness of a medication.[2] One benefit of the approach presented in this thesis is that it does not *require* longitudinal data, which is usually not available. However, that does not mean the approach could not be extended to include longitudinal information in future work.

Many previous studies have explored how language changes over time, in both healthy aging and in dementia. Much of the work examining language change over the lifetime has focused on writing, since few elderly people have speech samples from over the course of their lives. For example, Snowdon et al. (1996) analyzed the autobiographical writings of nuns, and found that certain linguistic measures could presage the diagnosis of Alzheimer's disease by decades. Heitkamp et al. (2016) also analyzed patient diaries, focusing on a single sv-PPA patient. They found linguistic changes 7 years prior to diagnosis, including changes to lexical

[2] There is currently no cure for dementia, although certain medications may provide symptomatic relief in some types of dementia (Herrmann et al., 2013).

diversity and word frequency, although frank semantic errors did not occur until later.

Another source of written data has been from novelists. Le et al. (2011) compared lexical and syntactic trends from three famous British novelists, and concluded that Agatha Christie's longitudinal pattern was more similar to that of Iris Murdoch (who died with Alzheimer's disease) than P.D. James, who was cognitively healthy at death. Van Velzen and Garrard (2008) analyzed the writing of Dutch author Gerard Reve and found a sharp decline in lexical diversity in his last novel, published shortly before his diagnosis with Alzheimer's disease.

In some cases, speech from public figures who are later diagnosed with dementia is available for retrospective analysis. Brian Butterworth published a study analyzing the speech patterns of U.S. President Ronald Reagan in the years preceding his eventual diagnosis of Alzheimer's disease (Erard, 2008). Butterworth reported that Reagan produced more sentence fragments and "slips of the tongue", paused more often, and spoke more slowly in 1984 relative to 1980. More recent work has used tools from natural language processing to add further evidence to this finding, showing a reduction in the number of unique words used, and an increase in non-specific nouns and fillers in Reagan's spontaneous speech (Berisha et al., 2015).

Longitudinal speech samples over shorter periods have also been analyzed as part of research studies on language in dementia. For PPA specifically, Bird et al. (2000) tracked three sv-PPA patients longitudinally over the course of the disease progression. They found that the participants produced more high-frequency, low-imageablity words and fewer nouns relative to verbs as the disease progressed. Thompson et al. (1997a) conducted a longitudinal analysis of narrative speech in four participants with nfv-PPA, and found two distinct patterns of decline: three participants resembled a set of participants with agrammatic Broca's aphasia, while one participant produced nonfluent but relatively grammatical speech up until 9 years post-onset. In later work, the authors suggested that this participant would today be classified as having logopenic PPA (Thompson et al., 2012). Kavé et al. (2007) analyzed narrative speech from an individual with sv-PPA over the course of three years, and found a severe decline in conceptual semantics, in marked contrast to her preserved morphological and syntactic abilities.

Likewise, longitudinal speech analysis has contributed to research on Alzheimer's disease. Ahmed et al. (2013) found that the majority of their participants showed subtle changes in narrative speech months before they were diagnosed with Alzheimer's disease, but that these differences were heterogeneous across the group, suggesting a potential benefit of comparing a particular individual's abilities longitudinally. Furthermore, these subtle changes may not have been detected by more coarse-grained language tasks. In contrast, Forbes-McKay et al. (2013) studied changes in Alzheimer's disease longitudinally, and calculated a number of different linguistic measures from picture description narratives collected 12 months apart. They found no significant difference on any of the measures except the number of phonological paraphasias. Using a computational approach, Yancheva et al. (2015) used features based on the work presented in this thesis to predict cognitive test scores over time in a cohort of Alzheimer's patients, to within a margin of error comparable to human inter-rater agreement.

In theory, automated analysis could make it possible to measure changes over much shorter time spans than have previously been studied, either via regularly scheduled assessments or continuous monitoring. The idea of continuously monitoring speech patterns has gained traction in the field of mental health in recent years. For example, Karam et al. (2014) developed a system to continuously and unobtrusively monitor cell phone conversations from users to recognize periods of mania and depression in bipolar disorder. This idea could also be applied to monitoring cognitive health. Kaye (2008) describes a smart-home environment that collects information about the resident's cognition through a speech-interface to the home's computer. One important application of such a system could be to track participants in clinical trials to detect conversion to mild cognitive impairment or Alzheimer's disease. Kaye (2008) lists some of the advantages over the current paradigm (participants travel to the study site to complete a neuropsychological battery every 6–12 months), including: ease of establishing an accurate baseline, ability to identify outlier "good days" and "bad days", increased accessibility for participants who have limited access to transportation or live in remote locations, and the potential for unobtrusive, natural measurements of day-to-day cognitive function.

Obviously, automated analysis of conversational speech is very different from the analysis of semi-structured narrative tasks that we consider here. Furthermore, there are ethical concerns related to continuous monitoring, largely relating to the privacy concerns of the individual being monitored and those with whom they interact. Individuals may also find it stressful or disturbing for their speech to be continually monitored. Longitudinal assessment in general would mark a departure from the work presented in this thesis in two major aspects: (1) An individual's language performance would be compared primarily against their own performance at a previous time, rather against group means, although comparison to normative data would still be useful to differentiate pathological change from normal changes due to aging. (2) Rather than using a classification framework to determine a binary distinction (healthy versus unhealthy), it would likely be more useful to use regression or time-series analysis to pinpoint an individual's continuous progression along a spectrum, in order to characterize their rate of decline and predict their future status from past measurements.

Indeed, the work presented here does not by itself immediately permit any of the applications outlined above. However, it may form part of the foundation on which these (and other) practical technologies may be built, to aid in the assessment and treatment of people with dementia.

Chapter 3

Classification of PPA from manual transcripts

This chapter focuses on the analysis of a set of oral narratives, collected as part of a study on primary progressive aphasia (PPA) in the Department of Speech-Language Pathology at the University of Toronto. The data set includes audio files of participants telling the story of Cinderella, along with manually-produced transcripts of the narratives. A full description of the participants, the elicitation task, and the process of transcription is given in the following section. Subsequent sections describe how we extracted features from the transcripts and audio files, and how we used those features to train classifiers to distinguish between participant groups. Several different experiments were performed.

Our first experiment is described in Section 3.2. Using features derived from the transcripts, we trained three classifiers to distinguish between participants with the semantic variant of PPA (sv-PPA) and controls, participants with the nonfluent/agrammatic variant of PPA (nfv-PPA) and controls, and between participants from the two subtypes. A second aim of that study was to compare our automatically-extracted features with the linguistic variables which were described in previous work on PPA and its subtypes. We found that features which were identified as important in our analysis had also been found to differentiate the groups in manual analyses of language in PPA.

Our next experiment, described in Section 3.3, involved adding acoustic features to the classification task. While we found the text features to be more discriminative, the acoustic features did hold some information and may be valuable when an accurate transcription is not available.

In Section 3.4 we present additional syntactic features that measure the frequency of production of different grammatical constituents. These fine-grained syntactic features were found to be more sensitive to the differences between sv-PPA and nfv-PPA than the syntactic complexity metrics we had originally considered.

Finally, in Section 3.5, we report the results of an ablation study to determine the optimal subset of features to use in each classification task. We were able to boost the best classification accuracy for each task by including only the most informative feature sets as input. Different feature sets were relevant to each classification task, reflecting the particular language impairments in the groups.

3.1 Introduction to the data[1]

3.1.1 Participants

Our participants comprised 24 patients diagnosed with either the semantic (sv-PPA) or non-fluent/agrammatic variant (nfv-PPA) of primary progressive aphasia (PPA), and 16 age- and education-matched healthy controls. The patient group is an unselected sample of patients with sv-PPA or nfv-PPA, except that participants who were unable to complete the narrative task ($n = 7$) were excluded: 2 nfv-PPA patients had incomprehensible speech, 1 nfv-PPA patient said nothing, 1 sv-PPA patient refused to attempt the task, and the responses of 1 nfv-PPA and 2 sv-PPA patients did not include any of the story that they were asked to tell, but instead

[1] The material presented in this section was previously published in: Kathleen C. Fraser, Jed A. Meltzer, Naida L. Graham, Carol Leonard, Graeme Hirst, Sandra E. Black, and Elizabeth Rochon (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex* 55, 43-60.

comprised statements of how they could not remember the story. Participants with PPA were recruited through three memory clinics in Toronto and each was diagnosed by an experienced behavioural neurologist. There were a further 7 patients in the cohort who were diagnosed with logopenic PPA, but this group was not included in the present study due to its small size. Control participants were recruited from a volunteer participant pool. All participants were native speakers of English, or completed some of their education in English. Exclusion criteria included a known history of drug or alcohol abuse, or a history of neurological or major psychiatric illness.

The study was approved by the Research Ethics Boards of all the hospitals involved in recruitment, as well as the board at the University of Toronto. Written informed consent was obtained from all participants.

Diagnosis was based on history, neuroimaging, neurological examination and neuropsychological testing, and all patients met current criteria for PPA (Gorno-Tempini et al., 2011). Patients with the fluent variant exhibited grammatically correct fluent speech, with word-finding difficulties. Those with the nonfluent variant had effortful, halting speech with anomia, although not all exhibited clear agrammatism in production or clear apraxia of speech on formal testing. Demographic data are listed in Table 3.1.

All participants underwent a battery of neuropsychological and linguistic tests as part of a longitudinal study of PPA which is being conducted in the Department of Speech-Language Pathology of the University of Toronto. The neuropsychological test information is reported in Table 3.1. The level of general cognitive functioning was measured using the Mini-Mental State Examination (Folstein et al., 1975) and the Dementia Rating Scale-R (Jurica et al., 2001). The two patient groups did not differ significantly on these tests, but were both impaired relative to controls. In keeping with the diagnosis of PPA, both patient groups performed poorly on a test of picture naming (Boston Naming Test, Kaplan et al., 2001) and on category fluency for animals (where participants are asked to name all the animals they can think of in 1 minute). Impairment in syntactic comprehension is a known feature of nfv-PPA and indeed was ex-

	sv-PPA ($n = 10$)	nfv-PPA ($n = 14$)	Controls ($n = 16$)	Group effect
Demographic information				
Age	65.6 (7.4)	64.9 (10.1)	67.8 (8.2)	ns
Years of education	17.5 (6.1)	14.3 (3.6)	16.8 (4.3)	ns
Sex	3 F	6 F	7 F	
Handedness	9 R	13 R	16 R	
General cognitive function				
Mini-Mental State Examination (/30)	24.4 (4.3)^{a}	25.0 (2.9)^{a}	29.3 (0.8)	***
Dementia Rating Scale-R (/144)	117.2 (12.6)^{a}	123.9 (15.6)^{a}	142.2 (1.7)	***
Language production				
Boston Naming (/60)	13.9 (7.3)^{a,b}	39.6 (11.5)^{a}	55.8 (3.3)	***
Category fluency — animals	7.6 (3.7)^{a}	12.3 (6.0)^{a}	20.4 (4.4)	***
Language comprehension				
Test for the Reception of Grammar (/80)	71.4 (11.0)	63.9 (12.0)^{a}	79.1 (0.9)	***
Peabody Picture Vocabulary Test (/204)	113.8 (30.8)^{a,b}	172.9 (14.3)^{a}	196.1 (3.9)	***
Visuospatial				
Copy of Rey Complex Figure (/36)	33.2 (2.6)	29.9 (5.3)^{a}	33.4 (1.4)	*
VOSP cube analysis subtest (/10)	9.4 (1.9)	9.2 (1.6)	8.5 (2.1)	ns
Nonverbal memory				
30 minute recall of Rey Complex Figure (/36)	12.7 (7.1)	14.9 (6.3)	18.8 (6.9)	ns
Nonverbal reasoning				
Raven's Coloured Progressive Matrices (/36)	31.5 (5.0)	27.1 (6.5)^{a}	31.8 (4.2)	*

Table 3.1: Demographic and neuropsychological data for each participant group. Values shown are mean (standard deviation). Asterisks denote significant effect of group on 1-way analyses of variance at * $p < .05$, *** $p < .001$.

[a] Significantly different from controls

[b] Significantly different from nonfluent patients

hibited by the nfv-PPA group studied here; this ability was measured using the Test for the Reception of Grammar (Bishop, 2003), and the nonfluent group (only) performed significantly worse than controls. Impairment in single-word comprehension is an established feature of sv-PPA; both of our patient groups were impaired on single-word comprehension, but as expected, the sv-PPA patients performed significantly worse than the nfv-PPA patients (Peabody Picture Vocabulary Test, Dunn and Dunn, 1997). Consistent with the diagnosis of PPA, performance was generally better on nonverbal tests. The nfv-PPA group was mildly impaired on copying the Rey Complex Figure (Rey, 1941) while the sv-PPA group showed normal performance. On another measure of visuospatial functioning, the cube analysis subtest from the Visual Object and Space Perception Battery (Warrington and James, 1991), both patient groups performed normally. Similarly, performance on nonverbal episodic memory was normal for both patient groups; this was assessed by asking participants to recall the Rey Complex Figure 30 minutes after copying it. Finally, nonverbal reasoning was relatively preserved, although it was mildly impaired for the nfv-PPA group (only) (Raven's Coloured Progressive Matrices, Raven, 1962).

3.1.2 Narrative task

Speech samples were elicited by having participants tell the Cinderella story, as in Saffran et al. (1989). To prompt their memories for the story, participants were given as much time as they needed to examine a picture book illustrating the story. When each participant had finished looking at the pictures, and the book had been removed, the examiner said "Now you tell me the story. Include as much detail as you can and try to use complete sentences." After letting the participant speak for as long as he or she wished, if the story was incomplete, general encouragement for more speech was given, for example, "Good, tell me more about that", "What happens next", "Go on", etc. At no time were specific questions or prompts given. The narratives were recorded on a digital audio recorder for subsequent verbatim transcription. Transcription was done in accordance with the procedures used in the Quantitative Production Analysis (Berndt et al., 2000), with the exception that punctuation and sentence initial capitalization were used. Brief pauses were marked with commas, while pauses longer than 1 second were timed and the length of the pause was noted (e.g., (2 sec)); however, commas and pauses were removed before analysis. Sentence boundaries were marked with full stops. Placement of sentence boundaries was guided by semantic, syntactic, and prosodic features, using a method essentially identical to that described by Thompson et al. (2012). When utterance boundaries were ambiguous, the segmentation which produced shorter utterances was preferred (as in Thompson et al. (2012) and Wilson et al. (2010)). Fillers such as *um* and *uh* were transcribed (and analyzed), but were not included in the total word count. Repetitions, false starts, and repeated but incomplete attempts at a given word were transcribed, but only repetitions of words and false starts were included in the total word count. Neologisms were transcribed with the International Phonetic Alphabet, and words/passages which were incomprehensible were marked with ###. Phonemic errors were written using the Roman alphabet and were followed by the transcriber's gloss of the word, which was put into double brackets. Neologisms and incomprehensible speech were not included in the automated analyses or in the word counts as we could not be certain how many words were represented; note, however, that incomprehensible passages were always brief. There were only rare instances of neologisms or incomprehensible speech: 3 participants (one in each group) had 2 occurrences each, while a further 3 participants had 1 occurrence (2 nfv-PPA, 1 sv-PPA).

3.2 Automatic classification using text features from manual transcripts[2]

This study had two aims. The first was to develop a machine learning classifier that would analyze speech samples and distinguish between control participants and participants with either nfv-PPA or sv-PPA, as well as between the two patient groups. The other aim of this study was

[2] The material presented in this section was previously published in: Kathleen C. Fraser, Jed A. Meltzer, Naida L. Graham, Carol Leonard, Graeme Hirst, Sandra E. Black, and Elizabeth Rochon (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex* 55, 43-60.

to identify the automatically extracted features that best distinguish the groups, and to compare this with results in the literature that are based on traditional (manual) analysis methods. Identification of the distinguishing features is important for improved detection and differentiation of the variants of PPA.

3.2.1 Features

The automatically extracted features are defined in Table 3.2. The first feature is the number of words in the transcript. The subsequent 22 structural features (2–23) were calculated using Lu's L2 Syntactic Complexity Analyzer (Lu, 2010), which uses the Stanford parser (Klein and Manning, 2003). We modified features 3–9 to be normalized by the total number of words, to facilitate comparison between narratives of different lengths. Lu used these features to analyze the syntactic complexity of college-level English essays from Chinese students, but the software has also been used to analyze spoken language (Chen and Zechner, 2011). We have not attempted to adapt this tool specifically for the study of aphasic speech; rather, we are interested to see how well such methods perform in the absence of any domain-adaptation.

The next four features (24–27) are also measures of syntactic complexity. Tree-based measures have been used to detect age-related cognitive decline (Cheung and Kemper, 1992) and mild cognitive impairment (Roark et al., 2011). Parse trees were constructed using the Stanford parser, so they are based on the same structural model as features 2–23. The Yngve depth quantifies to what extent the syntactic structure of a sentence contains left-branching rather than right-branching phrases, which provides a comparison metric of syntactic complexity (Yngve, 1960). For detailed illustrations of how Yngve depth is quantified, see Sampson (1997), Cheung and Kemper (1992), or Yngve (1960). We quantified Yngve depth as both mean depth over all words, the maximum depth in the sentence, and the total depth. For example, a sentence with an object-embedded relative clause such as *The juice that the child spilled stained the rug* is more left-branching than one with a subject-embedded relative clause such as *The child spilled the juice that stained the rug* (Stromswold et al., 1996). Using our procedures, the first

	Feature	Definition
1	Words	# words, excluding fillers, neologisms, and short, incomplete attempts at words
2	Sentences	# sentences, as marked by the transcriber
3	T-units	# T-units (a clause and all of its dependent clauses)
4	Clauses	# clauses (a structure consisting of at least a subject and a finite verb)
5	Coordinate phrases	# coordinate phrases (a phrase containing a coordinating conjunction)
6	Complex nominals	# complex nominals (a noun phrase, clause, or gerund that stands in for a noun)
7	Complex T-units	# complex T-units (a T-unit which contains a dependent clause)
8	Verb phrases	# verb phrases (a phrase consisting of at least a verb and its dependents)
9	Dependent clauses	# dependent clauses (a clause which could not form a sentence on its own)
10	Mean length of sentence	# words / # sentences
11	Mean length of clause	# words / # clauses
12	Mean length of T-unit	# words / # T-units
13	Dependent clauses per clause	# dependent clauses / # clauses
14	Dependent clauses per T-unit	# dependent clauses / # T-units
15	Verb phrases per T-unit	# verb phrases / # T-units
16	Clauses per sentence	# clauses / # sentences
17	Clauses per T-unit	# clauses / # T-units
18	Complex T-units per T-unit	# complex T-units / # T-units
19	Coordinate phrases per T-unit	# coordinate phrases / # T-units
20	Complex nominals per T-unit	# complex nominals / # T-units
21	T-units per sentence	# T-units / # sentences
22	Coordinate phrases per clause	# coordinate phrases / # clauses
23	Complex nominals per clause	# complex nominals / # clauses
24	Tree height	height of the parse tree
25	Total depth	total Yngve depth, averaged over all sentences
26	Max depth	maximum Yngve depth, averaged over all sentences
27	Mean depth	mean Yngve depth, averaged over all sentences
28	Nouns	# nouns / # words
29	Verbs	# verbs / # words
30	Noun-verb ratio	# nouns / # verbs
31	Noun ratio	# nouns / (# nouns + # verbs)
32	Inflected verbs	# inflected verbs / # verbs
33	Light verbs	# light verbs / # verbs
34	Determiners	# determiners / # words
35	Demonstratives	# demonstratives / # words
36	Prepositions	# prepositions / # words
37	Adjectives	# adjectives / # words
38	Adverbs	# adverbs / # words
39	Pronoun ratio	# pronouns / (# nouns + # pronouns)
40	Function words	# function words / # words
41	Frequency	mean frequency of all words appearing in the frequency norms
42	Noun frequency	mean frequency of nouns appearing in the frequency norms
43	Verb frequency	mean frequency of verbs appearing in the frequency norms
44	Imageability	mean imageability of all words appearing in the imageability norms
45	Noun imageability	mean imageability of nouns appearing in the imageability norms
46	Verb imageability	mean imageability of verbs appearing in the imageability norms
47	Age of acquisition	mean age of acquisition of all words appearing in the age of acquisition norms
48	Noun age of acquisition	mean age of acquisition of nouns appearing in the age of acquisition norms
49	Verb age of acquisition	mean age of acquisition of verbs appearing in the age of acquisition norms
50	Familiarity	mean familiarity of all words appearing in the familiarity norms
51	Noun familiarity	mean familiarity of nouns appearing in the familiarity norms
52	Verb familiarity	mean familiarity of verbs appearing in the familiarity norms
53	Type-token ratio	# unique word types / # words
54	Word length	mean number of letters in each word
55	Fillers	# fillers / # words
56	Um	# occurrences of ‘um’ / # words
57	Uh	# occurrences of ‘uh’ / # words
58	Speech rate	# words uttered / total time in minutes

Table 3.2: Definitions of features.

sentence is assigned these values: [max depth: 3, mean depth: 1.67, total depth: 15], while the second sentence is assigned: [max depth: 2, mean depth: 1.11, total depth: 10].

Features 28–40 rely explicitly on the part-of-speech (POS) tags for each word in the sample, determined by using the Stanford POS tagger (Toutanova et al., 2003). Differences in the noun and verb production of sv-PPA and nfv-PPA patients have been noted before (Harciarek and Kertesz, 2011). It has also been observed that nfv-PPA patients are more likely to omit inflections and function words (Harciarek and Kertesz, 2011). Here, function words included determiners, pronouns, prepositions, conjunctions, particles, and modals.

Word-level phonemic errors can present a potential problem to the tagger. We have used an additional tag called *not in dictionary* or *NID* for cases in which the speaker produces a nonword token. This prevents such tokens from being counted towards the wrong part-of-speech category. If a word error results in the substitution of another English word, this is not detected by our system and could be tagged incorrectly. However, the context around the word may provide useful clues to the tagger in such cases.

Verbs can be categorized as being *heavy* or *light*, according to their semantic complexity. Light verbs like *have* or *do* can be used in such a wide variety of different contexts that they are similar in some ways to closed-class function words (Breedin et al., 1998). We used the same list of light verbs as Breedin et al. (1998), namely: *be*, *have*, *come*, *go*, *give*, *take*, *make*, *do*, *get*, *move*, and *put*.

Features 41–52 measure frequency, imageability, age of acquisition, and familiarity. Frequency was calculated according to the SUBTL norms (Brysbaert and New, 2009), and the remaining three according to the combined Bristol norms and Gilhooly-Logie norms (Stadthagen-Gonzalez and Davis, 2006; Gilhooly and Logie, 1980). In addition to calculating the overall averages, the measures were calculated for nouns and verbs independently, to explore any possible dissociations. We calculated the proportion of words covered by the norms based on unique word forms (as opposed to individual occurrences). The coverage for the frequency norms is between .92 and .95 across the three groups. The coverage for the imageability, age of acquisition, and familiarity norms is lower, ranging from .25 to .31 for all content words across the three groups. One reason for this is that the Bristol norms were specifically designed to exclude high-frequency words, as the authors wanted to use words in "the frequency range most often sampled by psycholinguistic experiments" (Stadthagen-Gonzalez and Davis, 2006). This means that most of the words included in the norms have frequencies between 1 and 100 counts per million. For example, nouns like *thing* (which has a frequency of 1088 counts per million in the SUBTL norms) and *name* (641 counts per million) are excluded, as are verbs like *go* (3793 counts per million) and *do* (6135 counts per million).

The last six features (53–58) are measures of fluency and vocabulary richness. One way to measure vocabulary size is by calculating the type-token ratio, which is the ratio of the number of word types to the total number of words in the sample. A type-token ratio of 1.0 would mean that every word in the sample was unique; a low type-token ratio would indicate that many words were repeated. Filled pauses are measured by counting occurrences of the words *um*, *uh*, *ah*, and *er*, called "fillers" in Table 3.2. The words *um* and *uh* were also counted individually, because of research which suggests that they may be used to indicate major and minor pauses, respectively (Clark and Fox Tree, 2002). Finally, speech rate has been shown to distinguish between PPA patients and controls, although the results in sv-PPA are inconsistent. Here we consider an estimate of the speech rate, which was calculated by dividing the number of words produced by the participant by the total speech sample time.

Impaired speech can present difficulties for automatic language processing techniques, which have typically been developed for well-formed, written text. However, these methods represent a starting point for future development of more sophisticated techniques. To increase the probability of the structural analyses producing accurate results, our system counts and then removes the filled pauses from the transcript. Short nonword tokens (i.e., repeated but incomplete attempts at a given word, e.g., *br-* bring) are also removed.

To test the results of our automatic methods against traditional manual methods, we had a human annotator perform part-of-speech tagging and calculate a subset of the parse measures

	Clauses	Dependent clauses	T-units	Complex T-units	Coordinate phrases
Per narrative	0.9966	0.9319	0.9269	0.8792	0.9475
Per sentence	0.9630	0.6396	0.4756	0.4568	0.7921

Table 3.3: Correlations between human- and computer-generated counts for syntactic structures.

for three randomly-chosen narratives from each of the three participant groups (22.5% of the total data set). For part-of-speech tagging, we measured the agreement between the human annotator and the automatic tagger by counting the number of tags on which they both agreed, and dividing by the total number of tags. The average agreement was 87.3% for the nfv-PPA group, 89.2% for the sv-PPA group, and 91.9% for the control group. For comparison, the best reported accuracies for statistical taggers on written, well-formed text are around 97% (Manning, 2011), while Pakhomov et al. (2010a) reported a tagging accuracy of 86% on nfv-PPA speech transcripts.

In comparing the parse features, we were limited by the fact that Lu's program simply outputs counts for each measure, rather than the actual constituents being measured. This makes it impractical to use traditional parse measures such as the PARSEVAL measures (Manning and Schütze, 1999). Instead we had the annotator produce counts of different structures, just as the software does, and then correlated the two sets of counts (Pearson's correlation). We examined a set of features that seemed to be important in distinguishing the groups based on preliminary analyses, although not all of them were significant in the final version of the system. We measured the correlation in two ways: the correlation between the counts for each individual sentence, and the correlation between the total counts for each narrative. The correlation coefficients for these measures are shown in Table 3.3.

Correlations between the scores calculated by the human annotator and Lu's system were high when considered across narratives, which is the most relevant comparison for our purposes, as only the total scores for the narratives were used as input to the classifiers. The per-sentence correlations were somewhat lower. Inspection of the discrepancies between the manual and automatic scoring revealed a systematic pattern. In general, the automatic system has high agreement with the human annotator when determining the number of clauses. However, it has difficulty labelling clauses as being either independent or dependent, especially when the clauses are not connected by a conjunction. In many such cases, we found that the system counted a different number of dependent clauses than the human annotator, which in turn affected the number of T-units and complex T-units. Given this apparently systematic error, the results from Lu's syntactic complexity analyzer must be interpreted with caution.

3.2.2 Classification

We trained machine learning classifiers to predict participants' diagnoses based on a set of features extracted automatically from speech transcripts. Including too many features risks overfitting the classifier to idiosyncrasies in the training set, resulting in poor generalization to new data points. Therefore, some process of feature selection is necessary. To select features on which to train the classifiers, we conducted a two-tailed t-test on each feature between the two groups that were to be distinguished. All features that were significant at $p < 0.05$ were used for classification. The values of the selected features make up a feature vector, which defines a point in feature space. The goal of a machine learning classifier is to take a feature vector as input, and output a class label (in this case, either sv-PPA, nfv-PPA, or control). Three machine learning classifiers from the WEKA machine learning toolkit were compared (Hall et al., 2009): naïve Bayes, logistic regression, and support vector machines (SVM).

The classifiers were evaluated on the basis of classification accuracy, or the total proportion of narratives which were correctly classified. The evaluation was performed using leave-one-out cross-validation. In this procedure, one data point is left out, and the classifier is trained on the remaining data. The left-out data point can then be used as an unbiased test point. This procedure is repeated until each data point has been left out once, and the performance is averaged.

Feature		sv-PPA	Controls	*p* value	
1	Words	380.300 (272.429)	403.688 (121.380)	0.8025	
2	Sentences	20.400 (13.550)	15.688 (6.877)	0.3276	
3	T-units	0.069 (0.022)	0.058 (0.019)	0.2036	
4	Clauses	0.145 (0.013)	0.133 (0.011)	0.0292	*
5	Coordinate phrases	0.028 (0.010)	0.029 (0.007)	0.6676	
6	Complex nominals	0.098 (0.039)	0.083 (0.020)	0.2735	
7	Complex T-units	0.037 (0.009)	0.030 (0.009)	0.0505	
8	Verb phrases	0.177 (0.020)	0.167 (0.015)	0.2056	
9	Dependent clauses	0.067 (0.016)	0.052 (0.021)	0.0543	
10	Mean length of sentence	20.135 (8.190)	28.608 (10.973)	0.0346	*
11	Mean length of clause	7.035 (0.603)	7.602 (0.632)	0.0329	*
12	Mean length of T-unit	16.217 (6.481)	19.465 (8.042)	0.2701	
13	Dependent clauses per clause	0.460 (0.090)	0.391 (0.144)	0.1459	
14	Dependent clauses per T-unit	1.094 (0.497)	1.081 (0.768)	0.9571	
15	Verb phrases per T-unit	2.848 (1.276)	3.186 (1.196)	0.5094	
16	Clauses per sentence	2.845 (1.010)	3.738 (1.242)	0.0573	
17	Clauses per T-unit	2.292 (0.786)	2.558 (1.007)	0.4610	
18	Complex T-units per T-unit	0.580 (0.183)	0.535 (0.159)	0.5353	
19	Coordinate phrases per T-unit	0.463 (0.334)	0.589 (0.369)	0.3820	
20	Complex nominals per T-unit	1.558 (0.720)	1.694 (1.082)	0.7038	
21	T-units per Sentence	1.250 (0.230)	1.526 (0.356)	0.0249	*
22	Coordinate phrases per clause	0.193 (0.074)	0.221 (0.058)	0.3150	
23	Complex nominals per clause	0.673 (0.229)	0.627 (0.162)	0.5911	
24	Tree height	13.167 (2.359)	14.821 (2.401)	0.0998	
25	Total Yngve depth	70.477 (41.393)	117.609 (72.438)	0.0456	*
26	Maximum Yngve depth	5.091 (1.168)	6.023 (1.156)	0.0613	
27	Mean Yngve Depth	2.913 (0.409)	3.345 (0.498)	0.0250	*
28	Nouns	0.141 (0.031)	0.179 (0.026)	0.0051	**
29	Verbs	0.207 (0.028)	0.200 (0.019)	0.4427	
30	Noun-verb ratio	0.699 (0.209)	0.907 (0.163)	0.0164	*
31	Noun ratio	0.403 (0.072)	0.472 (0.047)	0.0178	*
32	Inflected verbs	0.635 (0.127)	0.706 (0.086)	0.1417	
33	Light verbs	0.474 (0.093)	0.476 (0.085)	0.9527	
34	Determiners	0.107 (0.030)	0.120 (0.016)	0.2203	
35	Demonstratives	0.037 (0.011)	0.012 (0.009)	0.0000	**
36	Prepositions	0.103 (0.035)	0.087 (0.015)	0.1994	
37	Adjectives	0.034 (0.013)	0.038 (0.009)	0.3434	
38	Adverbs	0.083 (0.017)	0.058 (0.014)	0.0010	**
39	Pronoun ratio	0.508 (0.094)	0.416 (0.068)	0.0175	*
40	Function words	0.467 (0.033)	0.453 (0.033)	0.2823	
41	Frequency	5.021 (0.105)	4.803 (0.104)	0.0001	**
42	Noun frequency	3.861 (0.231)	3.282 (0.183)	0.0000	**
43	Verb frequency	4.614 (0.282)	4.378 (0.184)	0.0341	*
44	Imageability	477.721 (44.025)	507.025 (20.643)	0.0729	
45	Noun imageability	560.959 (43.450)	580.710 (12.370)	0.1913	
46	Verb imageability	416.117 (37.506)	385.947 (22.543)	0.0387	*
47	Age of acquisition	258.881 (21.629)	257.814 (12.476)	0.8894	
48	Noun age of acquisition	254.246 (33.465)	251.696 (19.006)	0.8295	
49	Verb age of acquisition	260.465 (27.603)	266.521 (14.566)	0.5338	
50	Familiarity	607.358 (12.903)	565.956 (10.052)	0.0000	**
51	Noun familiarity	604.910 (20.954)	545.967 (17.119)	0.0000	**
52	Verb familiarity	605.526 (19.573)	600.218 (13.388)	0.4629	
53	Type-token ratio	0.405 (0.118)	0.415 (0.057)	0.8028	
54	Mean word length	3.735 (0.186)	3.997 (0.152)	0.0017	**
55	Fillers	0.053 (0.067)	0.054 (0.056)	0.9876	
56	Um	0.007 (0.008)	0.014 (0.015)	0.1613	
57	Uh	0.046 (0.061)	0.040 (0.060)	0.8052	
58	Speech rate	104.048 (35.149)	160.779 (35.131)	0.0007	**

Table 3.4: A comparison of sv-PPA and control features. Values shown are mean (standard deviation). Asterisks denote significance (* $p < .05$; ** $p < .01$).

Feature		nfv-PPA	Controls	*p* value	
1	Words	302.214 (141.837)	403.688 (121.380)	0.0466	*
2	Sentences	16.143 (12.526)	15.688 (6.877)	0.9049	
3	T-units	0.061 (0.023)	0.058 (0.019)	0.7698	
4	Clauses	0.141 (0.020)	0.133 (0.011)	0.2027	
5	Coordinate phrases	0.028 (0.013)	0.029 (0.007)	0.7585	
6	Complex nominals	0.092 (0.017)	0.083 (0.020)	0.2005	
7	Complex T-units	0.030 (0.010)	0.030 (0.009)	0.9345	
8	Verb phrases	0.167 (0.017)	0.167 (0.015)	0.9699	
9	Dependent clauses	0.055 (0.016)	0.052 (0.021)	0.7312	
10	Mean length of sentence	24.501 (12.192)	28.608 (10.973)	0.3437	
11	Mean length of clause	7.299 (0.986)	7.602 (0.632)	0.3348	
12	Mean length of T-unit	19.655 (8.814)	19.465 (8.042)	0.9516	
13	Dependent clauses per clause	0.384 (0.082)	0.391 (0.144)	0.1333	
14	Dependent clauses per T-unit	1.054 (0.545)	1.081 (0.768)	0.9106	
15	Verb phrases per T-unit	3.212 (1.400)	3.186 (1.196)	0.9577	
16	Clauses per sentence	3.346 (1.493)	3.738 (1.242)	0.4444	
17	Clauses per T-unit	2.716 (1.243)	2.558 (1.007)	0.7078	
18	Complex T-units per T-unit	0.517 (0.146)	0.535 (0.159)	0.7456	
19	Coordinate phrases per T-unit	0.615 (0.521)	0.589 (0.369)	0.8763	
20	Complex nominals per T-unit	1.767 (0.790)	1.694 (1.082)	0.8331	
21	T-units per Sentence	1.256 (0.254)	1.526 (0.356)	0.0232	*
22	Coordinate phrases per clause	0.203 (0.093)	0.221 (0.058)	0.5336	
23	Complex nominals per clause	0.660 (0.130)	0.627 (0.162)	0.5493	
24	Tree height	12.933 (2.718)	14.821 (2.401)	0.0555	
25	Total Yngve depth	108.405 (78.010)	117.609 (72.438)	0.7415	
26	Maximum Yngve depth	5.275 (1.364)	6.023 (1.156)	0.1197	
27	Mean Yngve Depth	3.085 (0.589)	3.345 (0.498)	0.2070	
28	Nouns	0.155 (0.040)	0.179 (0.026)	0.0644	
29	Verbs	0.191 (0.025)	0.200 (0.019)	0.2804	
30	Noun-verb ratio	0.837 (0.286)	0.907 (0.163)	0.4276	
31	Noun ratio	0.444 (0.082)	0.472 (0.047)	0.2775	
32	Inflected verbs	0.706 (0.096)	0.706 (0.086)	0.9923	
33	Light verbs	0.538 (0.141)	0.476 (0.085)	0.1666	
34	Determiners	0.130 (0.032)	0.120 (0.016)	0.3183	
35	Demonstratives	0.026 (0.018)	0.012 (0.009)	0.0210	*
36	Prepositions	0.088 (0.032)	0.087 (0.015)	0.8884	
37	Adjectives	0.030 (0.017)	0.038 (0.009)	0.1219	
38	Adverbs	0.069 (0.030)	0.058 (0.014)	0.2253	
39	Pronoun ratio	0.476 (0.095)	0.416 (0.068)	0.0601	
40	Function words	0.478 (0.045)	0.453 (0.033)	0.0968	
41	Frequency	4.962 (0.118)	4.803 (0.104)	0.0006	**
42	Noun frequency	3.451 (0.317)	3.282 (0.183)	0.0936	
43	Verb frequency	4.608 (0.237)	4.378 (0.184)	0.0072	**
44	Imageability	509.119 (40.552)	507.025 (20.643)	0.8634	
45	Noun imageability	579.078 (23.669)	580.710 (12.370)	0.8192	
46	Verb imageability	404.073 (49.196)	385.947 (22.543)	0.2215	
47	Age of acquisition	245.879 (14.862)	257.814 (12.476)	0.0260	*
48	Noun age of acquisition	235.038 (20.937)	251.696 (19.006)	0.0316	*
49	Verb age of acquisition	264.400 (13.607)	266.521 (14.566)	0.6834	
50	Familiarity	573.625 (18.408)	565.956 (10.052)	0.1808	
51	Noun familiarity	561.110 (24.689)	545.967 (17.119)	0.0668	
52	Verb familiarity	589.838 (19.242)	600.218 (13.388)	0.1043	
53	Type-token ratio	0.421 (0.046)	0.415 (0.057)	0.7421	
54	Mean word length	3.769 (0.136)	3.997 (0.152)	0.0002	**
55	Fillers	0.083 (0.080)	0.054 (0.056)	0.2584	
56	Um	0.025 (0.027)	0.014 (0.015)	0.1948	
57	Uh	0.058 (0.084)	0.040 (0.060)	0.5937	
58	Speech rate	78.468 (27.978)	160.779 (35.131)	0.0000	**

Table 3.5: A comparison of nfv-PPA and control features. Values shown are mean (standard deviation). Asterisks denote significance (* $p < .05$; ** $p < .01$).

Feature		sv-PPA	nfv-PPA	*p* value	
1	Words	380.300 (272.429)	302.214 (141.837)	0.4223	
2	Sentences	20.400 (13.550)	16.143 (12.526)	0.4436	
3	T-units	0.069 (0.022)	0.061 (0.023)	0.3518	
4	Clauses	0.145 (0.013)	0.141 (0.020)	0.6028	
5	Coordinate phrases	0.028 (0.010)	0.028 (0.013)	0.9334	
6	Complex nominals	0.098 (0.039)	0.092 (0.017)	0.6376	
7	Complex T-units	0.037 (0.009)	0.030 (0.010)	0.0580	
8	Verb phrases	0.177 (0.020)	0.167 (0.017)	0.2393	
9	Dependent clauses	0.067 (0.016)	0.055 (0.016)	0.0805	
10	Mean length of sentence	20.135 (8.190)	24.501 (12.192)	0.3056	
11	Mean length of clause	7.035 (0.603)	7.299 (0.986)	0.4254	
12	Mean length of T-unit	16.217 (6.481)	19.655 (8.814)	0.2829	
13	Dependent clauses per clause	0.460 (0.090)	0.384 (0.082)	0.0472	*
14	Dependent clauses per T-unit	1.094 (0.497)	1.054 (0.545)	0.8509	
15	Verb phrases per T-unit	2.848 (1.276)	3.212 (1.400)	0.5161	
16	Clauses per sentence	2.845 (1.010)	3.346 (1.493)	0.3381	
17	Clauses per T-unit	2.292 (0.786)	2.716 (1.243)	0.3188	
18	Complex T-units per T-unit	0.580 (0.183)	0.517 (0.146)	0.3825	
19	Coordinate phrases per T-unit	0.463 (0.334)	0.615 (0.521)	0.3953	
20	Complex nominals per T-unit	1.558 (0.720)	1.767 (0.790)	0.5084	
21	T-units per Sentence	1.250 (0.230)	1.256 (0.254)	0.9541	
22	Coordinate phrases per clause	0.193 (0.074)	0.203 (0.093)	0.7648	
23	Complex nominals per clause	0.673 (0.229)	0.660 (0.130)	0.8714	
24	Tree height	13.167 (2.359)	12.933 (2.718)	0.8246	
25	Total Yngve depth	70.477 (41.393)	108.405 (78.010)	0.1386	
26	Maximum Yngve depth	5.091 (1.168)	5.275 (1.364)	0.7270	
27	Mean Yngve Depth	2.913 (0.409)	3.085 (0.589)	0.4071	
28	Nouns	0.141 (0.031)	0.155 (0.040)	0.3591	
29	Verbs	0.207 (0.028)	0.191 (0.025)	0.1440	
30	Noun-verb ratio	0.699 (0.209)	0.837 (0.286)	0.1844	
31	Noun ratio	0.403 (0.072)	0.444 (0.082)	0.2122	
32	Inflected verbs	0.635 (0.127)	0.706 (0.096)	0.1550	
33	Light verbs	0.474 (0.093)	0.538 (0.141)	0.1935	
34	Determiners	0.107 (0.030)	0.130 (0.032)	0.0862	
35	Demonstratives	0.037 (0.011)	0.026 (0.018)	0.0688	
36	Prepositions	0.103 (0.035)	0.088 (0.032)	0.3086	
37	Adjectives	0.034 (0.013)	0.030 (0.017)	0.5577	
38	Adverbs	0.083 (0.017)	0.069 (0.030)	0.1586	
39	Pronoun ratio	0.508 (0.094)	0.476 (0.095)	0.4313	
40	Function words	0.467 (0.033)	0.478 (0.045)	0.5254	
41	Frequency	5.021 (0.105)	4.962 (0.118)	0.2139	
42	Noun frequency	3.861 (0.231)	3.451 (0.317)	0.0014	**
43	Verb frequency	4.614 (0.282)	4.608 (0.237)	0.9557	
44	Imageability	477.721 (44.025)	509.119 (40.552)	0.0916	
45	Noun imageability	560.959 (43.450)	579.078 (23.669)	0.2527	
46	Verb imageability	416.117 (37.506)	404.073 (49.196)	0.5036	
47	Age of acquisition	258.881 (21.629)	245.879 (14.862)	0.1211	
48	Noun age of acquisition	254.246 (33.465)	235.038 (20.937)	0.1309	
49	Verb age of acquisition	260.465 (27.603)	264.400 (13.607)	0.6845	
50	Familiarity	607.358 (12.903)	573.625 (18.408)	0.0000	**
51	Noun familiarity	604.910 (20.954)	561.110 (24.689)	0.0001	**
52	Verb familiarity	605.526 (19.573)	589.838 (19.242)	0.0659	
53	Type-token ratio	0.405 (0.118)	0.421 (0.046)	0.6828	
54	Mean word length	3.735 (0.186)	3.769 (0.136)	0.6341	
55	Fillers	0.053 (0.067)	0.083 (0.080)	0.3290	
56	Um	0.007 (0.008)	0.025 (0.027)	0.0335	*
57	Uh	0.046 (0.061)	0.058 (0.084)	0.6864	
58	Speech rate	104.048 (35.149)	78.468 (27.978)	0.0736	

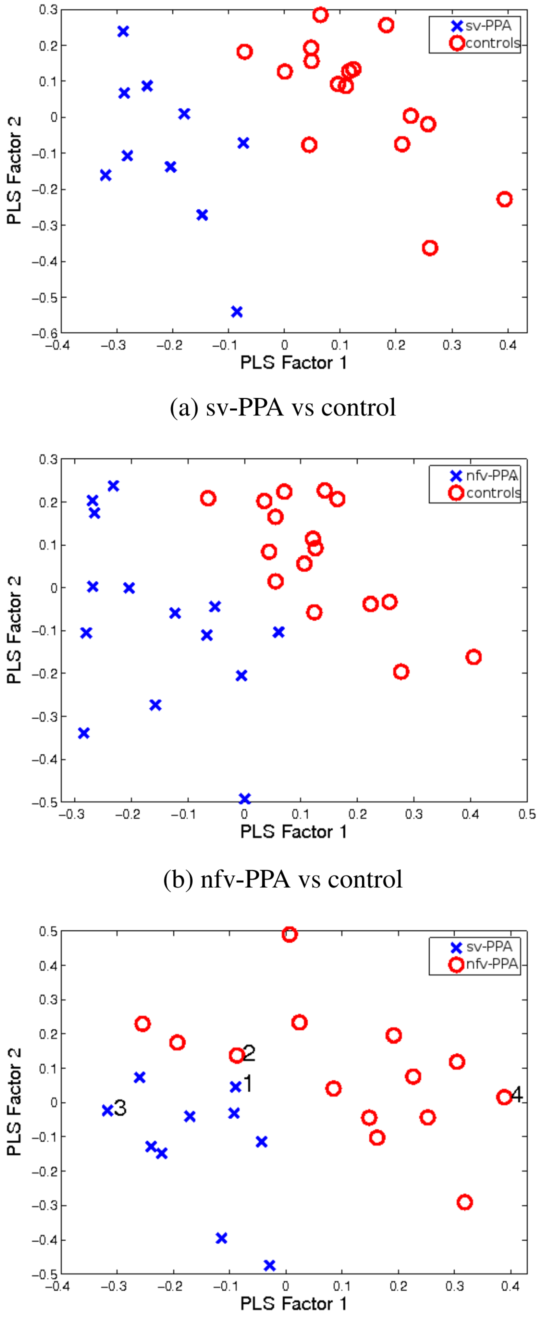
Table 3.6: A comparison of sv-PPA and nfv-PPA features. Values shown are mean (standard deviation). Asterisks denote significance (* $p < .05$; ** $p < .01$).

	sv-PPA vs. control	nfv-PPA vs. control	sv-PPA vs. nfv-PPA
Baseline	.615	.533	.583
Naïve Bayes	.923	.900	.792
Logistic Regression	.962	.933	.708
SVM	1.00	.967	.750

Table 3.7: Accuracies for the three classifiers, compared to a simple majority-class baseline classifier.

3.2.3 Results

We consider three separate classification tasks: (1) distinguishing between sv-PPA and controls; (2) distinguishing between nfv-PPA and controls; and (3) distinguishing between sv-PPA and nfv-PPA. The means and standard deviations for each attribute are compared in Tables 3.4 to 3.6. Group differences were measured using Welch's two-tailed, unpaired t-test, which does not assume that the two samples share the same variance. A significance level of $p < .05$ is indicated by a single asterisk, and $p < .01$ is indicated by a double asterisk. Because we were using the t-tests primarily as a means of feature selection, we did not adjust their significance levels for multiple comparisons. For each classification task, the set of significant features for that particular comparison formed the input vectors to the classifiers. The features were rescaled to have zero mean and unit variance before classification, to prevent features with large magnitudes (e.g., imageability) from dominating features with smaller magnitudes (e.g., fillers). The features that were considered significant between the sv-PPA and control transcripts, and therefore used in that classification task, were: number of clauses, mean length of sentence, mean length of clause, T-units per sentence, total Yngve depth, mean Yngve depth, nouns, noun-verb ratio, noun ratio, demonstratives, adverbs, pronoun ratio, frequency, noun frequency, verb frequency, verb imageability, familiarity, noun familiarity, mean word length, and speech rate. For the task of classifying nfv-PPA versus controls, the significant features were: number of words, T-units per sentence, demonstratives, frequency, verb frequency, age of acquisition, noun age of acquisition, mean word length, and speech rate. In the case of sv-PPA versus nfv-PPA, only five features were significant: dependent clauses per clause, noun

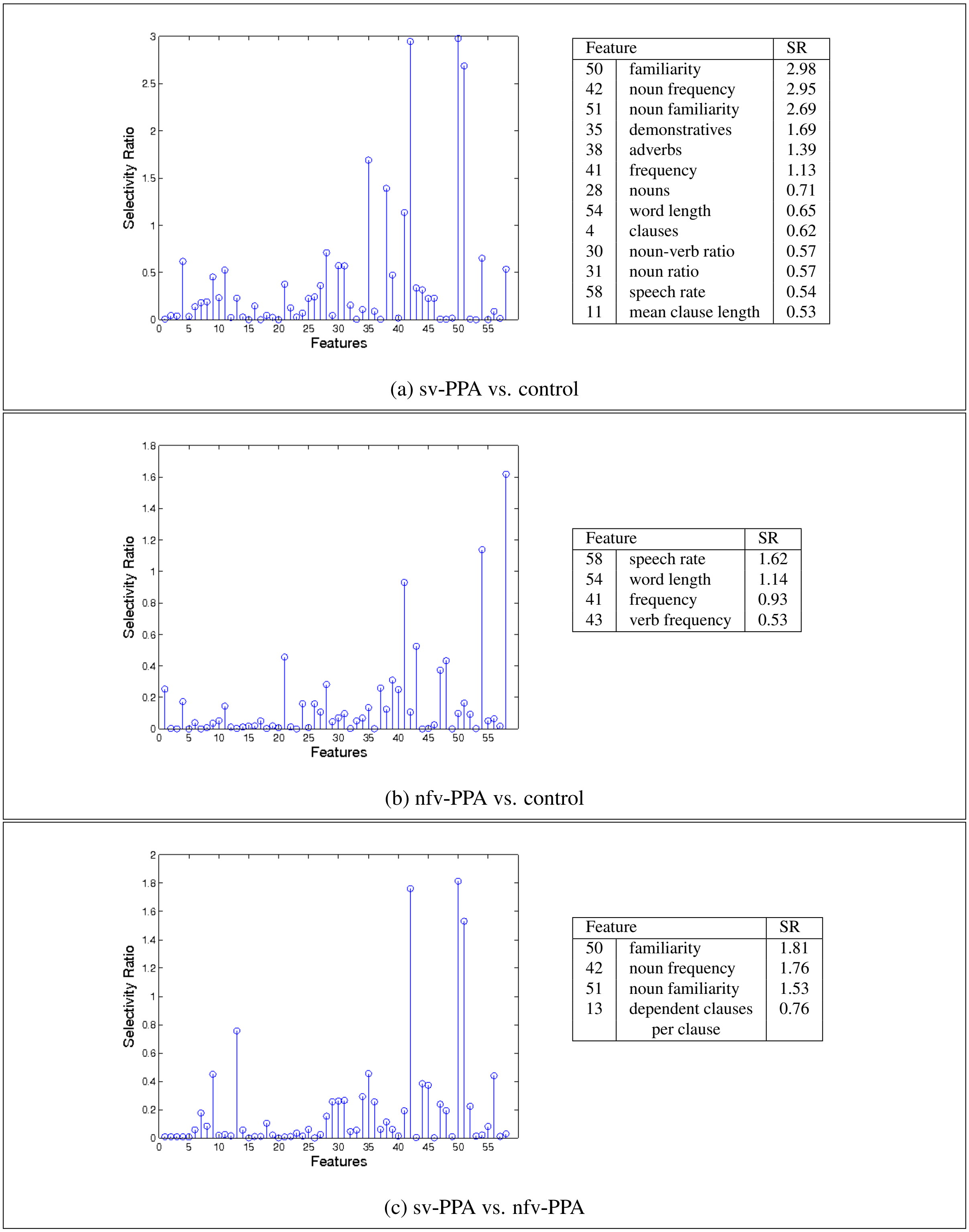


(a) sv-PPA vs control

(b) nfv-PPA vs control

(c) sv-PPA vs nfv-PPA

Figure 3.1: Partial least squares analysis of the data, with each point representing one transcript. Transcriptions for the participants labelled in Figure (c) are provided in Figures 3.3 to 3.6.



Feature		SR
50	familiarity	2.98
42	noun frequency	2.95
51	noun familiarity	2.69
35	demonstratives	1.69
38	adverbs	1.39
41	frequency	1.13
28	nouns	0.71
54	word length	0.65
4	clauses	0.62
30	noun-verb ratio	0.57
31	noun ratio	0.57
58	speech rate	0.54
11	mean clause length	0.53

(a) sv-PPA vs. control

Feature		SR
58	speech rate	1.62
54	word length	1.14
41	frequency	0.93
43	verb frequency	0.53

(b) nfv-PPA vs. control

Feature		SR
50	familiarity	1.81
42	noun frequency	1.76
51	noun familiarity	1.53
13	dependent clauses per clause	0.76

(c) sv-PPA vs. nfv-PPA

Figure 3.2: PLS selectivity ratio for each of the features. On the left, plots of the selectivity ratio for every feature. The feature numbers on the horizontal axes refer to the feature numbers in Table 3.2. Note the difference in scale on the vertical axes. On the right, the features with selectivity ratio greater than 0.5.

frequency, familiarity, noun familiarity, and occurrence of “um”.

The classification accuracies are given in Table 3.7. The baseline accuracies represent the accuracies that would be achieved by simply assigning every transcript to the larger of the two classes. That is, the baseline accuracy for sv-PPA ($n = 10$) versus controls ($n = 16$) would be achieved by simply classifying all the transcripts as controls ($16/26 = .615$). The two experimental scenarios involving patient groups versus control groups result in very high classification accuracies. The accuracies for classifying sv-PPA versus nfv-PPA transcripts are not as high; however, they are well above the baseline for all three of the classifiers.

For comparison, we also evaluated the performance of the classifiers trained on all features as input, rather than just the features pre-selected by the t-tests. This method was expected to perform rather poorly due to overfitting. Although using more features may improve classification on the training set, it results in poor generalization to new data, as assessed with the cross-validation procedure. Indeed, performance in this case was lower than the results shown in Table 3.7: for sv-PPA versus controls, the accuracies ranged from .846 to .923, for nfv-PPA versus controls the accuracies ranged from .700 to .800, and for sv-PPA versus nfv-PPA they ranged from .625 to .667. This illustrates the necessity of feature selection prior to classifier training.

Because the classification takes place in high-dimensional feature space, it is difficult to visualize the models produced by the classifiers. Instead, it is useful to visualize the classes in two dimensions by using some form of dimensionality reduction. Here we use the method of partial least squares, or PLS (Haenlein and Kaplan, 2004). PLS is similar to the well-known method of principal components analysis, except that principal components analysis discovers latent variables that best explain the variance in the attributes, while PLS discovers latent variables that are most predictive of the response (in this case, the patient groups or class labels). In addition, PLS is appropriate when the number of attributes is high compared to the number of data points, which is the case here. Scatter plots of the first two PLS components are shown in Figure 3.1.

Each of the plots shows relatively good separation between the groups. We were interested to see whether two narratives that were located close together in the PLS plot shared some similarities, even when the participants who produced them were from different diagnostic groups (for example, the points labelled 1 and 2 in Figure 3.1c). These transcripts are included in Figures 3.3 and 3.4. The transcripts associated with points 3 and 4 in Figure 3.1c are included in Figures 3.5 and 3.6 for the sake of comparison. Participants 1 and 2 have similar rates of speech, and Participant 1, although diagnosed with sv-PPA, makes several syntactic errors. Participant 2 was diagnosed with nfv-PPA, but tends to use high frequency words (such as *girls* instead of *stepsisters*). In contrast, the two transcripts from opposite sides of the PLS plots seem to be more clearly representative of their diagnostic groups.

> Well she was a kind of, a uh not a woman there, she was a bad woman. She had a couple of kids, women, girls, and he she had her too. And she uh, she's uh had to do all the work, do the washing and uh everyday she's at work and they don't have to work at all. And she used to go around tell her to do everything for them. So anyway uh [###] sort on it because uh then they came about this uh dance, and uh she wouldn't let her go. Two girls would go, but not her. But anyway, she got a fairy. Found a fairy a woman talked to her and that. And she got sitting up and they said, she got him to there with all dressed up to, uh I don't know the mice now, they helped her a lot apparently, but I don't know that too much about it, but they did. And she got real nice, she got dressed up she looked for every the stuff and got to the feet, the feet yeah the the shoes. And she went there with it, and when he went in, she uh uh what happened she drove ahead of it, and ran out, go away, and she dropped the (1) shoe. So anyway, this fellow came looking out, found you know, one of these fair fellows the big fellows, he went for the prince. [###] And he had to find that woman, I want to find that, and then he get this, he got the leg, the foot, the shoe. So he went over to everybody and when he found her, just when she found out and and and the prince took her away

Figure 3.3: Transcription 1 (sv-PPA, patient 29). Speech rate: 116.3 words per minute.

> OK. So, Cinderella, one day she ended up in, the middle of, um, in the in the in the, inside. So she was, she went in: to this house. And, sh: what basically what she had to do was she had to go down on her hands and knees and she had to clean. That was her job, um and, she also had, two or three, um, three girls, where they didn't have to do anything, which was interesting. And they were all excited about going to the ball. Um:, in the meantime, she, she also has a, grandmother, who looks af-, who looked after her. And then finally, she did actually, well first of all, she realized that the three people, they were her sisters as such, and then there was this other lady who basically was not a nice person at all. And: so that went on for quite a while, so they were, it was all about them and not about her, and then she actually met, her, beautiful person.

Figure 3.4: Transcription 2 (nfv-PPA, patient 41). Speech rate: 110.5 words per minute.

> Well, Cinderella, um (2) came to um, the co- n- I gon' say cottage but it's not a cottage obviously it's a mm quite a p- a place. And um, she uh initially is not th:ought of too much and uh the little animals sort of thing are b- uh down there and uh and as a matter of fact uh they uh come to like her as a matter of fact too. And um she uh is uh going there and these little things are going along and uh the uh older lady and so on uh not th- that g- good to her ah I I suspect. And then there's a shellow ((fellow)) who uh looks quite nice and he thinks that she looks quite nice and it ends up to some degree when they eventually get married. That's all I can think of right there, but sorry. (2) Uh (2) and there's all all s- uh sort of little little things that are running around and uh and that but part of the thing but she doesn't br- b- bother them but she rather likes that the gentleman who is probably one of the king's dau- s- s- sons. And away they go. Anyways that's (hhh) just looking at it slightly like that. I'm sure I missed a lot of things but nevertheless that's positive and I'm glad she's fi- having fun. And so are the (hhh) little things too. Anyways, there.

Figure 3.5: Transcription 3 (sv-PPA, patient 2). Speech rate: 101.5 words per minute.

> Okay. Um, Cinderella, they have uh a stepfathers and the uh godmother or whatever. Um (4) uh um (11) I, I guess they're going to go to the ball and then um all those mice they're going to do the dress (hhh) and um (3), and u:m oh the pumpkin (hhh) and uh (3) well the godmother will do the pumpkin and um (4), and then she'll go to the ball (3) and then (4) he has, he has a (3) him (hhh) him um anyways, so anyways twelve o'clock she'll go down the stairs. And she has a glass slipper, but it's, it's not it's just in the stairs, so anyways she um she had a pumpkin. She went (2) um went to her castle again and then um (3) and then they had a glass slipper for her and then: the prince and the Cinderella.

Figure 3.6: Transcription 4 (nfv-PPA, patient 31). Speech rate: 54.6 words per minute.

Rather than analyze the factor loadings, which are not easily interpretable in PLS, we calculate *selectivity ratios*, which are closely related to the correlation between each attribute and the response (Kvalheim, 2010). A high selectivity ratio indicates that a feature is very influential with respect to the response. The details of calculating this measure are given by Kvalheim (2010); we used a pre-existing Matlab package to perform the analysis (Li, 2011). Figure 3.2 shows the selectivity ratios for each feature in the three cases. For ease of interpretation, each feature with a selectivity ratio of greater than a cut-off of 0.5 is given for each of the three classification problems. In the majority of cases, these are the same features which were found to be significant and included in the classifiers above. We expect there to be some discrepancies, as the individual t-tests do not take into account correlations between variables, while the PLS analysis does. Selectivity ratios for influential features may be reduced in the presence of a second feature highly correlated with the first, as the two share variance that predicts group membership.

3.2.4 Discussion

In this study, we set out to determine whether computational methods could reliably distinguish between healthy controls and patients with PPA, as well as differentiate the two patient groups, based upon samples of narrative speech. We found that even with relatively short samples of narrative speech (i.e., for machine learning purposes), classifiers were able to achieve this goal with a high degree of accuracy. In addition, we wished to determine how the automatically extracted features compared to previous findings in the literature with respect to these two subtypes. In general, we found that our procedure identified many of the same features that have been previously noted to differ between groups (e.g., word frequency, speech rate, demonstrative pronouns), with some surprising findings (e.g., the lack of syntactic features as differentiating ones) and some new features identified (e.g., adverbs and word length). We discuss these issues below.

Classification

Our results show that machine learning classifiers can distinguish between controls and each of the two patient groups, sv-PPA and nfv-PPA, with a high degree of accuracy. Although less accurate than in comparison to controls, they also distinguish well between the two patient groups. The performance of the classifiers varied across the three tasks: SVM achieved the highest accuracy for sv-PPA versus controls and nfv-PPA versus controls, while naïve Bayes performed best for sv-PPA versus nfv-PPA. Logistic regression achieved the second-highest accuracy in the first two cases, but was the worst at distinguishing between sv-PPA and nfv-PPA. We also note the relatively high accuracy of naïve Bayes despite obvious correlations between the features in some cases.

Features that distinguished the groups and comparison with previous findings on PPA

The PLS analysis identified the features that best predicted group membership. The selectivity ratios, together with the group means on each feature, provided valuable information on the characteristics of narrative speech in each group.

The features that best distinguished the sv-PPA patients from controls were higher familiarity and frequency of words (particularly nouns), increased production of adverbs and demonstratives, production of shorter clauses, and reduced word length and speech rate. The familiarity of the words produced, which was greater in sv-PPA, was the feature with the highest selectivity ratio. The familiarity of nouns in particular was also ranked highly. The finding that the sv-PPA patients produced words with higher familiarity ratings than controls is consistent with findings from studies of naming (Lambon Ralph et al., 1998; Woollams et al., 2008), but to the best of our knowledge has not been previously documented in connected speech. It also fits with previous research demonstrating that sv-PPA patients' semantic knowledge is best preserved for familiar items (Funnel, 1996; Simons et al., 2001).

The sv-PPA patients tended to use higher frequency nouns than the controls, and this feature had the second highest selectivity ratio. Overall word frequency also distinguished well between the groups and these findings demonstrate the robust effect of frequency on language production in sv-PPA. It has been established that frequency has a pervasive influence on naming in sv-PPA (Lambon Ralph et al., 1998; Woollams et al., 2008), but this influence is less well documented in connected speech. Two studies which used picture description tasks observed that sv-PPA patients produced higher frequency nouns than controls (Bird et al., 2000; Wilson et al., 2010). Meteyard and Patterson (2009) did not evaluate frequency per se, but their analysis of structured interviews showed that patients with sv-PPA tended to replace content words with higher frequency (and less specific) words. We have found that patients with sv-PPA use higher frequency words overall, and that nouns are particularly affected.

Familiarity and frequency, the features with the highest selectivity ratios, are correlated with each other (Tanaka-Ishii and Terada, 2011). It is interesting to note that individually they may have had even higher selectivity ratios if only one of the features had been included in the model.

The sv-PPA patients produced more demonstratives and adverbs than controls, and this distinguished the groups. The increased reliance on demonstrative pronouns, which comprise the words *these*, *those*, *this*, *that*, *here*, and *there*, may reflect the tendency of sv-PPA patients to make substitutions of less specific words (Meteyard and Patterson, 2009) and use vague terms (Kavé et al., 2007; Patterson and MacDonald, 2006); the automated analysis techniques used here did not enable us to evaluate whether these terms had clear referents when used. Previous studies have documented over-reliance on pronouns in sv-PPA (Kavé et al., 2007; Meteyard and Patterson, 2009; Patterson and MacDonald, 2006; Wilson et al., 2010), but to the best of our knowledge this is the first examination specifically of demonstrative pronouns. The increased use of adverbs may at least partially reflect that fact that some participants in this patient group started many of the sentences in their narratives with *then* or *so*, sometimes preceded by *and*. Because *then* and *so* are classified as adverbs (describing when or why something happened), this inflates the count on these words. Repeated use of the same syntactic structure is compatible with the idea that sv-PPA patients are unable to produce the full range of syntactic structures (Benedet et al., 2006), but needs to be investigated further before firm conclusions can be drawn.

The sv-PPA patients produced fewer nouns than the controls, and reduced noun production accounted for 3 of the features which had high selectivity ratios (nouns, noun-to-verb ratio, and noun ratio). Difficulty with production of nouns is an expected finding. Impaired confrontation naming is a core diagnostic feature in sv-PPA (Gorno-Tempini et al., 2004), and is expected to lead to corresponding problems with word finding in connected speech (see Sajjadi et al., 2012). Moreover, previous studies have documented reduced production of nouns in the connected speech of people with sv-PPA (Bird et al., 2000; Ash et al., 2009; Patterson and MacDonald, 2006; Kavé et al., 2007). Bird et al. (2000) contrasted noun and verb production by sv-PPA patients on a picture description task and provided evidence that production of nouns was more affected as a result of their lower relative frequency. This is compatible with the current results in that our sv-PPA patients produced higher frequency, and proportionally fewer, nouns than controls (and these features had high selectivity ratios).

The mean word length for the sv-PPA patients was slightly shorter than for controls, and this feature distinguished well between the groups. The effect of word length has not been examined in connected speech. We attribute this small but significant effect of length to availability of words, rather than to difficulty with pronouncing long words. Many of the long words used by controls were used less often by the sv-PPA patients. The most frequently used long word for both groups was *Cinderella*, which was used a total of 69 times by the controls and a total of 17 times by the sv-PPA patients (an average of 4.3 versus 1.7 times per narrative). Other long words that were used more often by controls than patients include, for example, *slipper* which was used 48 times by controls but never by sv-PPA patients, and *beautiful* which was used 25 times by controls and 5 times by sv-PPA patients (an average of 1.6 versus 0.5 times per narrative).

The number, and mean length, of clauses both distinguished between the sv-PPA and control groups, although the numerical differences were rather small. The sv-PPA patients produced more clauses than the controls, but on average their clauses were shorter. The explanation for this is not clear. It seems unlikely to be an artifact of the automatic coding, as the clause counts had high agreement with the human annotator, even with the sentence-by-sentence comparison. It may be associated with the reduced production of nouns, which could result in fewer nouns per clause. In addition, inspection of the transcriptions indicates that the sv-PPA patients were more likely to produce “filler” comments such as *you know* and *whatever you call it*, which are relatively short clauses. We know of no identical analyses in other studies, although some have used similar measures: Patterson and MacDonald (2006) found that sv-PPA and controls produced similar numbers of clauses, while Sajjadi et al. (2012) found that sv-PPA patients produced fewer syntactically complex clauses relative to controls (data on simple clauses were not reported). Further work would be required to understand the basis and potential significance of the present finding that sv-PPA patients tend to produce shorter, but more, clauses than controls.

The final feature which distinguished the sv-PPA patients and controls was speech rate. Findings with respect to speech rate in sv-PPA have been inconsistent. The slower rate for sv-PPA patients may be due to pauses in speech while searching for a word, and seems unlikely to reflect a motor speech problem. Wilson et al. (2010) also documented slower speech rate in sv-PPA patients than in controls, but found that the maximum speech rate (which they defined as the three most rapidly spoken sequences of ten or more words) for their patients was normal; they suggested that the slower rate reflected impairment in higher-level processes, and we concur with this idea.

The features that distinguish sv-PPA patients from controls can largely be attributed to the semantic memory impairment. This seems to be the dominant influence in the language output of this group, and can account for the increased reliance on more familiar and frequent words, use of general terms such as demonstratives and adverbs like *then* and *so*, reduced production of nouns, and pauses for word-finding (which could explain the relatively slower speech rate).

The features that distinguished the nfv-PPA patients from controls were reduced speech rate and word length, as well as higher frequency of words and of verbs in particular. Not surprisingly, slower speech rate was the feature that best distinguished nfv-PPA from controls. A reduced rate of speech is one of the diagnostic features (Gorno-Tempini et al., 2004), and has been documented in other studies of connected speech production in nfv-PPA (Ash et al., 2010, 2006; Graham et al., 2004; Knibb et al., 2009; Thompson et al., 2012).

The nfv-PPA patients tended to produce shorter words than controls, and this feature had a high selectivity ratio. Increased word length of stimulus items has been shown to deleteriously affect naming, reading and repetition in nfv-PPA (Croot et al., 1998; Graham et al., 2004), and this effect could arise from phonological or motor speech impairment(s), which are common in nfv-PPA. In the case of narrative speech, it could also arise from word finding difficulties, which would affect availability of words (as suggested for the sv-PPA patients). The current analyses do not inform the choice of explanation, and indeed the use of shorter words could arise from different causes in different individuals.

Like the sv-PPA patients, the nfv-PPA patients were distinguished from controls by overall word frequency. For the nfv-PPA patients (vs. controls), verb frequency also had a high selectivity ratio. Once again we have a situation where two correlated features both have high selectivity ratios, suggesting that each would have had an even higher ratio if they had been included in the model without the other. Studies have shown that naming of verbs/actions in nfv-PPA is more impaired than naming of nouns/objects (Cotelli et al., 2006; Hillis et al., 2004). Although we do not know of a study which assesses the effect of frequency on naming of verbs in nfv-PPA, our results suggest that production of higher frequency verbs is better preserved. The finding that the nfv-PPA patients produce higher frequency verbs than the controls cannot be due to excessive use of light verbs (which tend to be high in frequency), because the t-tests indicate no difference between nfv-PPA and controls in use of light verbs. Despite this apparent difficulty with verb production, the nfv-PPA patients in this study produced nouns and verbs in normal proportions as demonstrated by equivalent noun-to-verb ratios for nfv-PPA patients and controls. An increase in noun-to-verb ratio is taken to indicate difficulty with verb production, and has been reported in other studies of connected speech in nfv-PPA (Thompson et al., 1997a). Consistent with the current findings, Graham et al. (2004) also documented normal noun-to-verb ratios but suggested that the verbs produced by nfv-PPA patients were less specific than those produced by controls.

The results also yielded interesting findings regarding the features that distinguished between the two patient groups. The sv-PPA patients used nouns which were more frequent and familiar, and words which were more familiar overall, than the nfv-PPA patients. These features are identical to those that best discriminated between sv-PPA patients and controls, and have the same rank order with respect to their selectivity ratios. As noted above, frequency and familiarity are known to affect naming performance in sv-PPA (Woollams et al., 2008; Lambon Ralph et al., 1998), and clearly these factors affect word finding in connected speech as well.

The greater impact (relative to nfv-PPA) of familiarity and frequency upon the speech of the sv-PPA patients may be due to their greater semantic impairment, or to the fact that they are more anomic than the nfv-PPA patients.

The final feature that differentiated the two patient groups was the number of dependent clauses per clause. This indicates that the sv-PPA patients produce a proportionally greater number of clauses which could not form sentences on their own. In their study of connected speech in sv-PPA, Meteyard and Patterson (2009) documented increased production of restarts, which they defined as a sentence which was incomplete and then started again. We did not count restarts, but inspection of the transcriptions reveals that there were many, and this could account at least partially for the increased production of incomplete (i.e., dependent) clauses. The finding that production of dependent clauses was greater in sv-PPA than nfv-PPA should, however, be regarded as tentative as there is uncertainty (noted above) in the ability of the system to properly identify dependent clauses. Further work would be needed to clarify the reliability and interpretation of this result.

Some surprising findings also emerged. While it is well documented that nfv-PPA patients tend to have sparse output, producing fewer words than either controls or sv-PPA patients (Graham et al., 2004; Wilson et al., 2010), total word count did not emerge as an important feature distinguishing the groups in the PLS plots. It is also surprising that so few of the features measuring syntactic complexity emerged as main distinguishing features, particularly for the nfv-PPA group. This may have occurred because not all of the nfv-PPA patients exhibited agrammatism. Alternatively, it may be due to the way the syntactic analyses were performed. As mentioned in Section 3.2.1, Lu's syntactic complexity analyzer was originally designed to be applied to written documents by second-language learners, and so may not be ideally suited to the analysis of aphasic speech. Our goal was to test its performance in this domain, and although we found that it was effective in detecting clause boundaries, and returned counts that were highly correlated with manual counts, it did encounter particular difficulty in labelling clauses as dependent or independent. In part, this may be attributable to the uncertainty about sentence boundaries as determined by a human transcriber, as opposed to quantitative linguistic criteria. We note that sentence boundaries are frequently ambiguous in natural speech, aphasic or otherwise. Further methodological development of automated analysis for syntactic complexity is a promising avenue for future research, particularly if it can be made to operate mainly within rather than across clause boundaries.

To summarize, the automated analyses indicated that sv-PPA patients showed an over-reliance on words which were high in familiarity and/or frequency, and this applied particularly to nouns. They also produced proportionally fewer nouns, but more demonstratives (e.g., *this*, *these*) and more adverbs (e.g., *so*, *then*). In contrast, the speech of the nfv-PPA patients was characterized by reduced speech rate and word length; this group also tended to use words which were high in frequency and this applied particularly to verbs. Verbs were, however, produced in normal proportions. The sv-PPA patients were distinguished from nfv-PPA by their relatively greater use of words with higher familiarity and frequency.

3.3 Augmenting text features with acoustic features[3]

The work described in the previous section used textual features extracted from transcripts of speech to classify between sv-PPA, nfv-PPA, and healthy controls. The classifiers achieved high accuracies between patient groups and controls, although the accuracies were reduced when attempting to distinguish between the two PPA subtypes. In this section, we analyze *acoustic* features of patient and control speech, and augment the text-based classifiers with these features. A benefit to acoustic features is that they do not require transcription, as they can be calculated directly from the audio files.

[3] The material presented in this section was previously published in: Kathleen Fraser, Frank Rudzicz, and Elizabeth Rochon (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Proceedings of INTERSPEECH*, pp.2177–2181.

3.3.1 Features

Slow, effortful speech is one of the core symptoms of nfv-PPA, and apraxia of speech can be an early feature, including the production of speech sound errors and disordered prosody (Gorno-Tempini et al., 2011; Grossman, 2010). Amici et al. (2006) reported over 80% of their nfv-PPA patients had apraxia of speech. More broadly, atypical $F0$ range and variance have been shown to be indicative of articulatory neuropathologies within the context of speech recognition (Mengistu et al., 2011; Kent and Kim, 2003). In contrast, speech production is generally spared in sv-PPA, although sv-PPA patients may produce long pauses as they search for words (Wilson et al., 2010).

We attempt to measure some of these characteristics by introducing a set of acoustic features, given in Table 3.8. We follow the work of Pakhomov et al. (2010b) and measure pause-to-word ratio (i.e., the ratio of silent segments longer than 150 ms to non-silent segments), mean fundamental frequency ($F0$) and variance, total duration of speech, long pause ($>$ 0.4 ms) count, and short pause ($>$ 0.15 ms and $<$ 0.4 ms) count. To this we add mean pause duration and phonation rate (the amount of the recording spent in voiced speech) (Roark et al., 2011), as well as the mean and variance for the first 3 formants ($F1, F2, F3$) and mean instantaneous power, which we expect to relate to dysprosody (Baghai-Ravary and Beet, 2012), mean and maximum first autocorrelation function (Meilán et al., 2014), skewness and kurtosis of the signal (Lee and Hahn, 2010), zero-crossing rate, mean recurrence period density entropy (a method for measuring the periodicity of a signal, which has been applied to pathological speech generally (Little et al., 2006)), jitter (Silva et al., 2009), and shimmer (Adnène et al., 2003).

In addition to the acoustic features, we consider the same text features as before (see Table 3.2). We extract 58 lexical and syntactic features from the transcript and an additional 23 acoustic features from the audio file, for a total of 81 possible features.

Total duration of speech Total length of all non-silent segments, in milliseconds.

Phonation rate Total duration of active speech divided by the total duration of the sample (including pauses).

Mean pause duration Mean length of pauses > 150 ms.

Short pause count Number of pauses > 150 ms and < 400 ms.

Long pause count Number of pauses ≥ 400 ms.

Pause:word ratio Ratio of silent segments longer than 150 ms to non-silent segments.

Mean/var. F0:3 Mean and variance of the fundamental frequency and first three formant frequencies.

Jitter Measure of the short-term variation in the pitch (frequency) of a voice.

Shimmer Measure of the short-term variation in the loudness (amplitude) of a voice.

Zero-crossing rate (ZCR) An approximation for average pitch of an utterance, defined as the number of sign changes along a signal, per second.

Mean instantaneous power Measure related to the loudness of the voice.

First autocorrelation function Mean and maximum of the first autocorrelation function.

Skewness Measure of lack of symmetry in the distribution of the amplitude of a signal, associated with a tense or "creaky" voice.

Kurtosis Measure of the "peakedness" of a signal's amplitude, or specifically the 4th moment of its distribution.

Mean recurrence period density entropy (MRPDE) Measure of periodicity of a signal. Specifically, it measures the extent to which a time series repeats itself. It is similar to linear autocorrelation.

Table 3.8: Acoustic features extracted directly from the PPA speech samples.

3.3.2 Feature selection

To avoid overfitting, we reduce the dimensionality of our data to be bounded above by the minimum number of data points available for a classification task; since there are 24 speakers with either nfv-PPA or sv-PPA, we reduce our feature space from 81 to at most 20. We compare two methods of performing this feature selection. In the first, similar to the method used in the previous section, we calculate Welch's t-test for each feature to calculate the significance of the difference in that feature between the two groups. We then rank each feature by its p-value, and for a feature set of size n we consider only the top n features from the ranked list. This method does not take into account any correlations between variables, but it does offer some insight into which individual features are most strongly indicative of one diagnosis (class) or the other. Feature selection methods based on p-value have been used in previous studies on machine learning classification of frontotemporal lobar degeneration (Peintner et al., 2008) and mild cognitive impairment (Roark et al., 2011).

The second method we consider is minimum-redundancy-maximum-relevance (mRMR) feature selection in which a set of features is selected such that redundancy (i.e., the average mutual information between features) is minimized and the relevance (i.e., the mutual information between the given features and the class) is maximized (Peng et al., 2005). Specifically, for feature f_i in set S and class c, mRMR selects the feature set S^* such that

$$S^* = \arg\max_{S} \left[\frac{1}{\|S\|} \sum_{f_i \in S} I(f_i;c) - \frac{1}{\|S\|^2} \sum_{f_i, f_j \in S} I(f_i;f_j) \right],$$

where $I(X;Y)$ is the mutual information between X and Y.

Table 3.9 shows the top $n = 10$ features selected by both the mRMR and p-value methods. Both mRMR and the p-value method select more textual features than acoustic features from all available features; 7/10 features are textual in all cases except for the p-value selection of features for the nfv-PPA versus sv-PPA case, in which case 9/10 of the selected features are textual. This might be tempered to some extent by the fact that our acoustic features are more highly correlated (average $r = 0.16\,(\sigma = 0.47)$ among acoustic features and $r = 0.05\,(\sigma = 0.37)$ among textual features). In general, the mRMR and *p*-value methods are in greater agreement on the PPA versus control task (with 6, 7, and 6 features in common across feature sets) than on the nfv-PPA versus sv-PPA task (with 5, 4, and 4 features in common). Also, the PPA and control classes are more significantly differentiated by the associated top $n = 10$ features than the nfv-PPA and sv-PPA classes; all features selected across feature sets in the PPA versus control case are significant at $\alpha \leq 0.05$ (mean $p = 0.003, (\sigma = 0.008)$) but fewer than half of the features across feature sets in the nfv-PPA versus sv-PPA case are significant (mean $p = 0.131, (\sigma = 0.156)$).

3.3.3 Results

Our experiments compare diagnostic accuracy across a number of empirical variables, namely the task (PPA vs. controls or sv-PPA vs. nfv-PPA), feature set ('Feat. set': text-only, acoustic-only, all), classifier (naïve Bayes (NB), support vector machine (SVM), and random forests (RF)), number of features considered for classification ('Num. feat.': 2, 5, 10, 15, 20), and the method of feature selection used to derive these reduced sets ('Feat. select', described in Section 3.3.2). We optimize the SVM classifier over several combinations of kernel (polynomial of degree 1 or 2, or radial basis function) and complexity ($c = \{0.01, 0.1, 1, 10, 100\}$). We optimize RF for the number of trees ($I = \{5, 10, 15, 20\}$) and the random seed ($S = \{1, 2, 3, 4, 5\}$). Accuracies of diagnostic classification were obtained for each of the possible permutations of our empirical parameters using leave-one-out cross-validation.

Table 3.10 shows the results of a multi-way analysis of variance (ANOVA) across each of our empirical variables. Each empirical parameter contributes significantly to the variance *except* for the number of features. Table 3.11 partitions rates of accurate diagnosis across tasks, feature sets, classifiers, and methods of feature selection. As expected, distinguishing PPA participants from controls is significantly easier than classifying among sub-types of PPA ($p < 0.0001$). Interestingly, although the effect size of considering *all* possible features rather than

		mRMR	p-value
PPA versus controls	text	***frequency***, ***verbalRate***, totalDepth, nounFrequency, verbFrequency, verbs, aveLengthWord, *demonstratives*, totalWords, *familiarity*	verbalRate‡, frequency‡, aveLengthWord‡, demonstratives‡, nounFamiliarity‡, nounFrequency‡, familiarity‡, verbFrequency‡, nouns‡, pronounRatio†
	acoustic	***phonationRate***, *meanF2*, *meanRPDE*, *skewness* pause:wordRatio, meanDurationOfPauses, meanInstantaneousPower, F0variance, longPauseCount, meanF0	phonationRate‡, meanRPDE‡, longPauseCount‡, short- PauseCount‡, meanDurationOfPauses‡, meanInstantaneousPower‡, shimmer‡, skewness†, kurtosis*, pauseWordRatio*
	all	***phonationRate***, familiarity, F2variance, *verbalRate*, nounFrequency, verbFrequency, *meanRPDE*, *nounRatio*, TUnitsPerSentence, demonstratives	phonationRate‡, verbalRate‡, frequency‡, aveLengthWord‡, demonstratives‡, meanRPDE‡, nounFamiliarity‡, longPauseCount‡, nounFrequency‡, familiarity‡
nfv-PPA versus sv-PPA	text	nounFamiliarity, *verbalRate*, ***imageability***, *nounFrequency*, adjectives, ***familiarity***, determiners, dependentClauses, nounAOA, S	familiarity‡, nounFamiliarity‡, nounFrequency‡, dependentClausesPerClause*, um*, complexTUnits, dependentClauses, verbFamiliarity, demonstratives, determiners
	acoustic	***ZCR***, ***shortPauseCount***, *skewness*, totalDurationOfSpeech, *F0variance*, *jitter*, shimmer, meanF0, meanRPDE, phonationRate	meanFirstAutocorrFunc*, jitter, totalDurationOfSpeech, maxFirstAutocorrFunc, pauseWordRatio, F3variance, F2variance, meanF2, meanF0, longPauseCount
	all	*imageability*, ***familiarity***, **jitter**, *verbalRate*, *nounFrequency*, clausesPerTUnit, nounFamiliarity, shortPauseCount, ZCR, demonstratives	familiarity‡, nounFamiliarity‡, nounFrequency‡, dependentClausesPerClause*, um*, meanFirstAutocorrFunc*, complexTUnits, dependentClauses, verbFamiliarity, demonstratives

Table 3.9: Selected features ($n = 10$) for each task and feature set using the mRMR and p-value methods. Features in **bold** and *italic* appear in the associated selected feature sets for $n = 2$ and $n = 5$ of the mRMR method, respectively; features marked with ‡, † and * represent features on which the given classes are significantly different at $\alpha = 0.005$, $\alpha = 0.01$, and $\alpha = 0.05$, respectively.

only textual features is small (Cohen's $d = 0.1363$), the difference in accuracy is significant ($p < 0.05$). If we consider each task separately, adding acoustics to textual features always increases accuracy, but not significantly (PPA vs. controls: $\mu_{text} = 90.4\%$, $\mu_{all} = 91.2\%$, $p =$

	Mean sq.	F	p
Task	7132.24	196.26	$2.27e-32$
Feat. select.	679.69	18.70	$2.31e-05$
Feat. set	2700.66	74.31	$1.96e-25$
Num. feat.	50.59	1.39	0.24
Classifier	985.84	27.13	$2.88e-11$

Table 3.10: Multi-way ANOVA (F statistics and p values) on accuracy across task, feature selection method, feature set, number of features, and classifier. Statistically significant ($\alpha = 0.05$) results are in **bold**.

0.24; sv-PPA vs. nfv-PPA: $\mu_{text} = 78.8\%$, $\mu_{all} = 80.4\%$, $p = 0.17$).

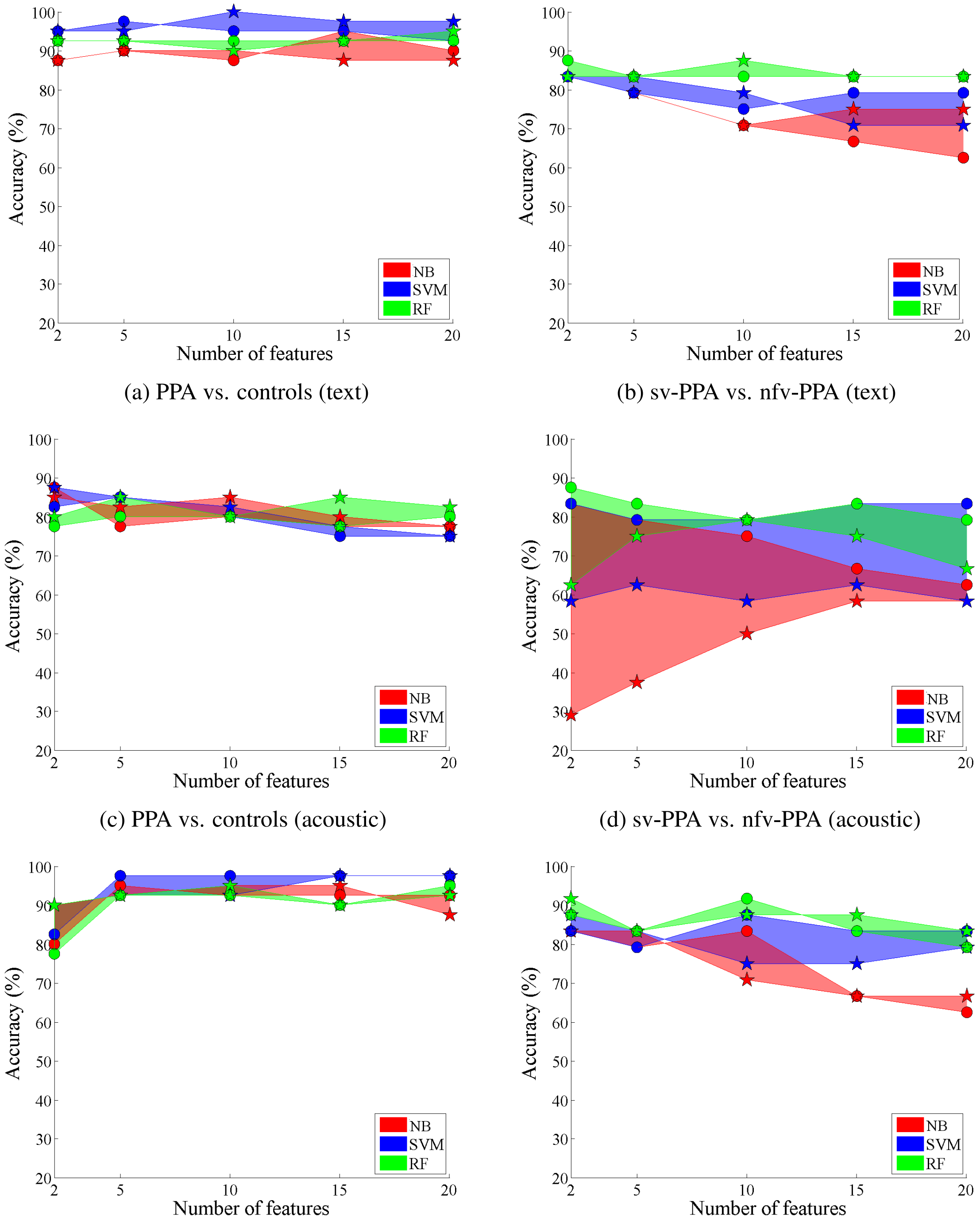
Variable	Value	Accuracy (%)
Task	PPA vs controls	$\mu = 87.39, (\sigma = 6.79)$
	sv-PPA vs nfv-PPA	$\mu = 75.59, (\sigma = 11.37)$
Feat. set	text	$\mu = 84.62, (\sigma = 8.51)$
	acoustic	$\mu = 74.05, (\sigma = 11.59)$
	all	$\mu = 85.80, (\sigma = 8.84)$
Classifier	NB	$\mu = 77.48, (\sigma = 12.87)$
	SVM	$\mu = 82.13, (\sigma = 11.30)$
	RF	$\mu = 84.85, (\sigma = 6.95)$
Feat. select.	pvalue	$\mu = 83.73, (\sigma = 8.28)$
	mRMR	$\mu = 80.37, (\sigma = 12.08)$

Table 3.11: Average accuracy μ (and standard deviation σ) across experiments partitioned by task, feature set, classifier, and method of feature selection.

Figure 3.7 shows graphs of accuracy over the number of features in feature sets for each type of feature set (text, acoustic, all) and each task (PPA vs. control and sv-PPA vs. nfv-PPA).

3.3.4 Discussion

The naïve Bayes method performed surprisingly well here, although generative approaches can sometimes outperform discriminative ones on small data sets (Ng and Jordan, 2002). The feature selection method can affect the accuracy of NB, as illustrated in Figure 3.7d. For a feature set of size two, the NB classifier achieved an accuracy of 83% using the p-value



(a) PPA vs. controls (text)

(b) sv-PPA vs. nfv-PPA (text)

(c) PPA vs. controls (acoustic)

(d) sv-PPA vs. nfv-PPA (acoustic)

(e) PPA vs. controls (all)

(f) sv-PPA vs. nfv-PPA (all)

Figure 3.7: Accuracies across task (PPA vs. control, sv-PPA vs. nfv-PPA) and feature set (text, acoustic, all) for NB (red), SVM (blue), and RF (green). Star-shaped points represent accuracies obtained using mRMR; circles represent those obtained using the *p*-value method.

method, and only 29% using mRMR. We hypothesize that this is because the *p*-value filter, which chooses features on the basis of their mean and variance, is choosing exactly those features which would best distinguish the groups in a Gaussian framework. Indeed, the top two features chosen by mRMR (ZCR and short pause count) both have bimodal (non-Gaussian) distributions.

The utility of the acoustic features appears to vary depending on the classification task. For the task of classifying PPA versus controls, it is unsurprising that features like phonation rate and short and long pause counts are informative, reflecting the language difficulties experienced by the PPA patients. Other features which were significantly different between the groups, including mean RPDE, mean instantaneous power, shimmer, skewness, and kurtosis, have not to our knowledge been previously reported for PPA.

However, only one acoustic feature, mean first autocorrelation function, significantly differentiated the sv-PPA and nfv-PPA groups. This lack of distinguishing features is somewhat unexpected, as sv-PPA patients are often described as more "fluent" than nfv-PPA patients, suggesting they could be distinguished on the basis of characteristics such as phonation rate and number of pauses. However, fluency may be a poor marker for subtyping PPA, in part because patients without nfv-PPA may show "intermittent dysfluency" due to word-finding difficulties (Thompson et al., 2012). Wilson et al. (2010) also found reduced average speech rate for both nfv-PPA and sv-PPA, and suggested that measuring maximum speech rate might be more useful for distinguishing them. In general, although acoustic features have a practical advantage in that they can be extracted directly from the audio file, their usefulness appears to be limited to the case of differentiating between PPA patients and controls.

3.4 New syntactic features for PPA classification

In our initial analysis of PPA using the manual transcripts (Section 3.2), we found that only one syntactic feature differed significantly between sv-PPA and nfv-PPA (dependent clauses per clause, $p = 0.047$), and between nfv-PPA and controls (T-units per sentence, $p = 0.023$). This was somewhat surprising, since the literature suggested that nfv-PPA speech would exhibit reduced syntactic complexity and increased agrammatism. However, it is not clear whether our results indicate a lack of syntactic impairment in the nfv-PPA group in our study, or whether the syntactic metrics we used were simply not sensitive enough.

Motivated by this uncertainty, in this section we present a new set of features that measure the frequency of different context-free grammar constituents in each narrative sample. The context-free grammar (CFG) is a popular formalism for modelling the structure of language, based on the idea that language can be described by a set of rules (also called constituents or productions) and a set of words and symbols (also called the lexicon) (Jurafsky and Martin, 2000). By embedding these rules hierarchically, we can construct sentences of arbitrary complexity. An example of a CFG representation of a sentence is given in Figure 3.8. By counting the frequency of occurrence of each production, we can compare the frequency distribution of different grammatical constructions across groups.

To test that these features could detect atypical syntax, we first conducted a study comparing Cinderella narratives from patients with post-stroke agrammatic aphasia with healthy controls (Fraser et al., 2014b). We found that we could achieve a better classification accuracy using the CFG features than traditional measures of syntactic complexity, and that many of the features could be interpreted with respect to previous findings on agrammatic speech, such as a reduced number of prepositional phrases, difficulty with grammatical morphemes (e.g., possessive markers), and an increase in filled pauses and repetitions. Given this successful proof-of-concept, we now examine whether these features uncover syntactic differences between the PPA and control groups, and between the two PPA subtypes.

Syntactic complexity metrics derived from parse trees have been used by various researchers in studies of mild cognitive impairment (Roark et al., 2011), autism (Prud'hommeaux et al., 2011), and child language development (Sagae et al., 2005; Hassanali et al., 2013). Here we focus specifically on the use of CFG production rules as features. Using the CFG production

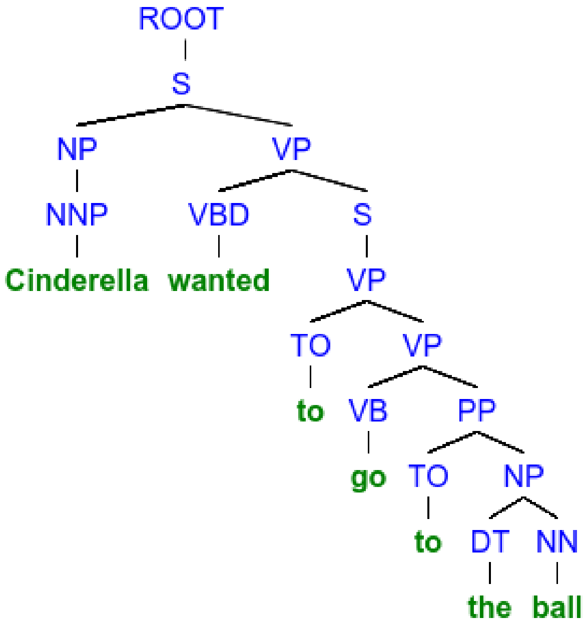


Figure 3.8: Context-free grammar (CFG) representation of a sentence. We count the frequency of occurrence of each non-lexical production, e.g., ROOT → S, S → NP VP, NP → NNP, and so on. We do not count productions containing terminal symbols (words), e.g., NNP → Cinderella.

rules from statistical parsers as features was first proposed by Baayen et al. (1996), who applied the features to an authorship attribution task. More recently, similar features have been widely used in native language identification (Wong and Dras, 2011; Brooke and Hirst, 2012; Swanson and Charniak, 2012). Perhaps most relevant to the task at hand, CFG productions as well as other parse outputs have proved useful for judging the grammaticality and fluency of sentences. For example, Wong and Dras (2010) used CFG productions to classify sentences from an artificial error corpus as being either grammatical or ungrammatical. Taking a different approach, Chae and Nenkova (2009) calculated several surface features based on the output of a parser, such as the length and relative proportion of different phrase types. They used these features to distinguish between human and machine translations, and to determine which of a pair of translations was the more fluent.

To our knowledge there has been no work using CFG features to assess the grammaticality of speech from individuals with dementia or aphasia. In related work, Meteyard et al. (2014)

counted the frequency of some syntactic constructions in sv-PPA and controls; they found an absence of frank agrammatism but reported a reduced range of syntactic constructions in sv-PPA, as well as some evidence for syntactic simplification.

3.4.1 Feature extraction

For the CFG production rules, we use the Charniak parser (Charniak, 2000) to parse each utterance in the transcript and then extract the set of non-lexical productions. Our choice of parser is based on the recommendations of Wong and Dras (2010). The total number of production types is large, many of them occurring very infrequently, so we compile a list of the 100 most frequently occurring productions in each of the three groups (sv-PPA, nfv-PPA, and controls) and use the combined set as the set of features. Here we consider integer frequency counts for each feature. The CFG non-terminal symbols follow the Penn Treebank naming conventions (Santorini, 1990).

3.4.2 Analysis of differences between groups

Rather than looking at individual features, of which there are 134 in our analysis, we first consider the distributions across features. To make the task more manageable, we group the features by the left-hand side of the rule. In this way, we can examine whether there are differences between the groups in how they choose to construct different types of phrases. Here we restrict our analysis to noun phrases, verb phrases, and prepositional phrases.

Figure 3.9 shows the frequency counts for different noun phrase (NP) structures in each group, Figure 3.10 shows the frequency counts for different verb phrase (VP) structures, and Figure 3.11 shows the frequency counts for different prepositional phrase (PP) structures. Raw frequency counts are used rather than ratios, to allow for χ^2 testing, below.

Looking at the bar graphs, we are able to pick out some qualitative differences between the groups. For example, when we consider the noun phrases in Figure 3.9, we observe lower counts for NP $\rightarrow$ NNP in the two PPA groups relative to controls, and greater counts for

NP → DT (especially considering the smaller size of the PPA groups). To examine such differences in a more rigorous way, we perform statistical testing to determine (1) whether there is a significant difference between the distributions, and (2) which features are contributing to the difference.

To compare the distributions of categorical data, we use the χ^2 test. This statistical test can be used to measure whether two (or more) population distributions are identical (McHugh, 2013). The null hypothesis is that the data were all drawn from the same distribution. Here, the null hypothesis is that patients from each group do not significantly differ in their preference for each grammatical construction. One benefit of χ^2 analysis is that it is appropriate for comparing unequal sample sizes, as we do here (McHugh, 2013). A limitation of this formulation is that χ^2 analysis assumes that the data were independently sampled, when clearly a number of the constituents will have been produced by the same person, and therefore are not independent. However, this assumption is often violated in practice (McHugh, 2013).

For each constituent type, our analysis is the same: first perform a χ^2 test to determine whether there is a significant difference between the groups. If so, then perform a post-hoc χ^2 test (with Bonferroni adjustment) to determine the effect between each pairwise combination of groups. The results of these tests are in Table 3.12. If a significant effect is found, we determine the individual variables which contribute to the lack of fit between the distributions. Following Agresti (2003), we use the Pearson standardized residual to identify features of interest. This method requires a threshold value; Agresti (2003) suggests a cut-off of “about 2 or 3” (p. 81). We split the difference and use a threshold of 2.5. Those features with a standardized residual greater than 2.5 are reported in Table 3.13.

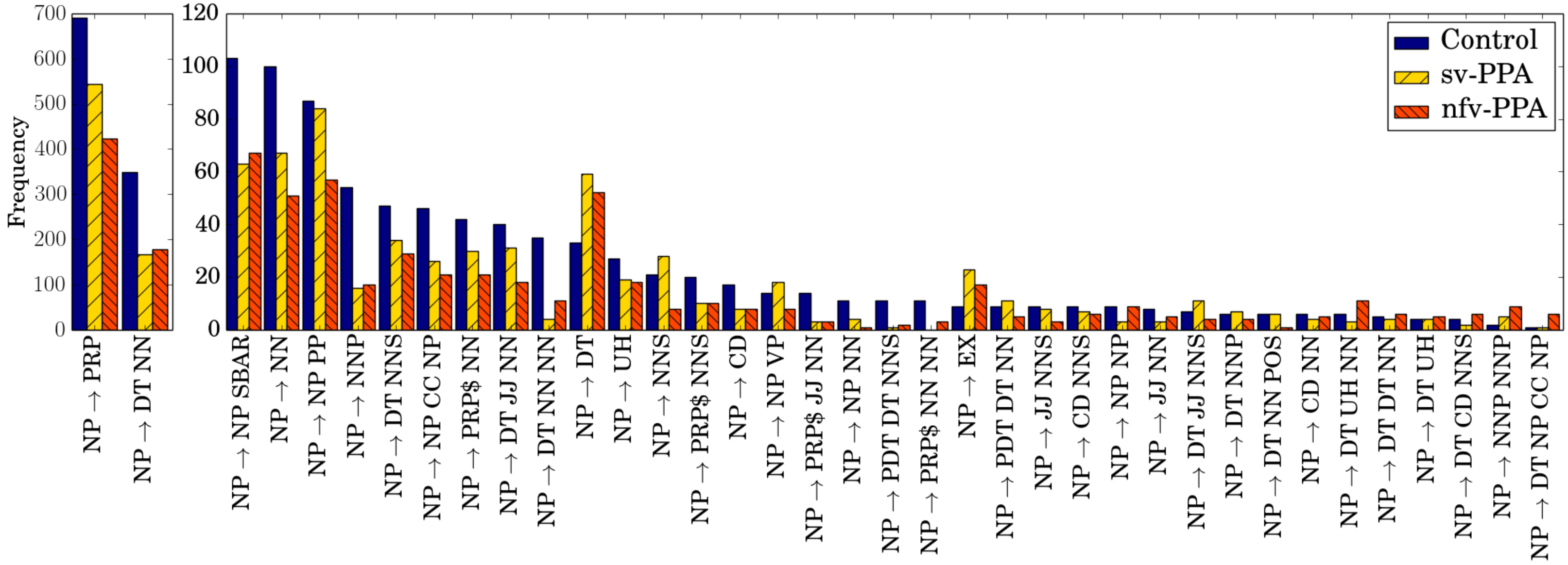


Figure 3.9: Frequency counts of noun phrase productions in each group.

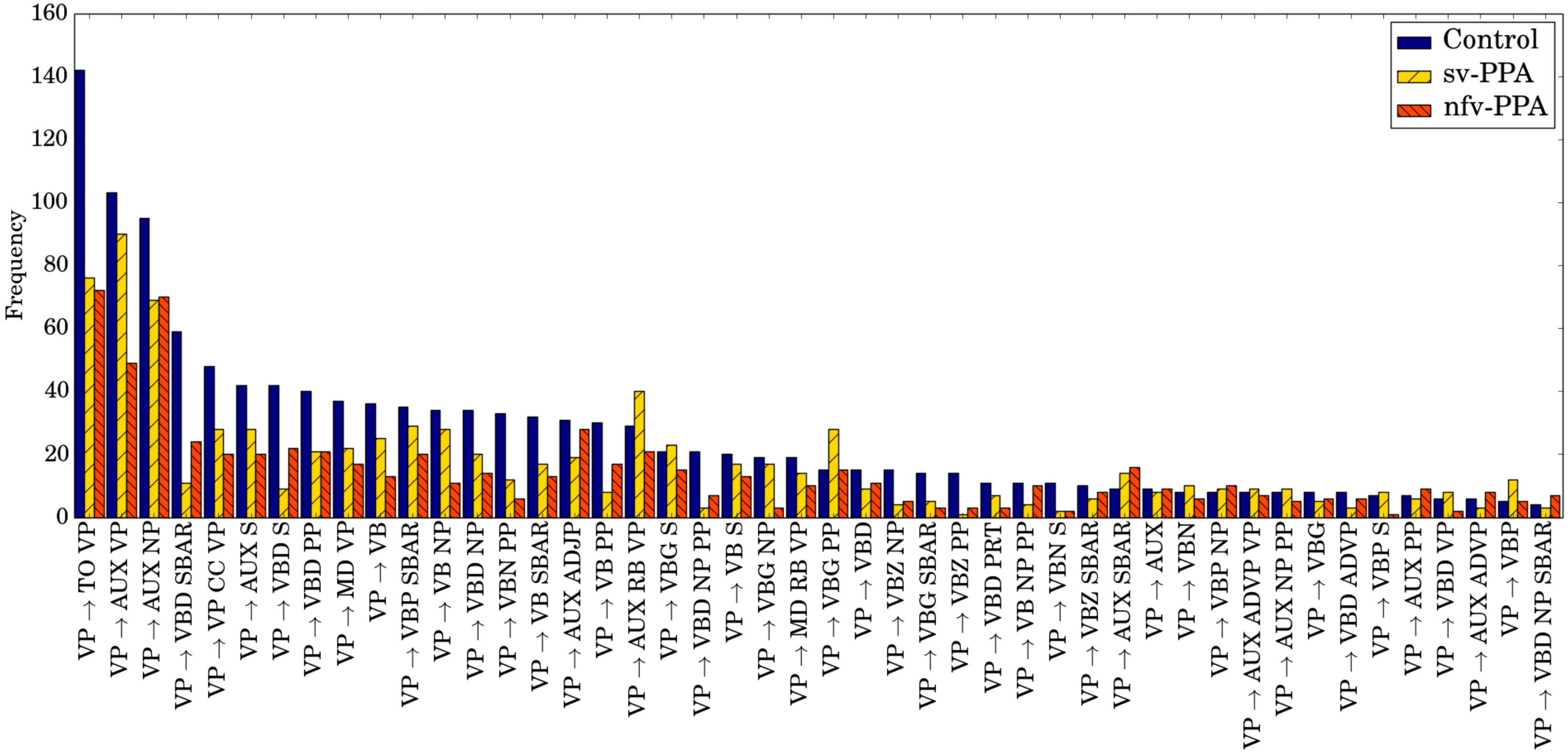


Figure 3.10: Frequency counts of verb phrase productions in each group.

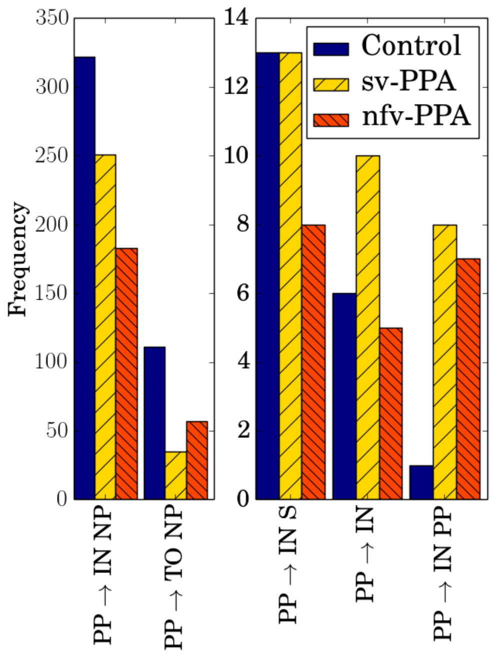


Figure 3.11: Frequency counts of prepositional phrase productions in each group.

Constituent head			χ^2	d.f.	p
VP			185.9471	90	1.192e-08*
	Post-hoc:	controls vs. sv-PPA	115.5328	45	4.045e-08*
		controls vs. nfv-PPA	75.0373	45	0.003283*
		sv-PPA vs. nfv-PPA	85.8254	45	0.0002352*
NP			192.8707	72	5.555e-13*
	Post-hoc:	controls vs. sv-PPA	124.5016	36	1.12e-11*
		controls vs. nfv-PPA	89.2601	36	2.04e-06*
		sv-PPA vs. nfv-PPA	63.7956	36	0.002913*
PP			33.378	8	5.264e-05*
	Post hoc:	controls vs. sv-PPA	31.7741	4	2.128e-06*
		controls vs. nfv-PPA	9.8792	4	0.04251
		sv-PPA vs. nfv-PPA	13.3384	4	0.009735*

Table 3.12: Results of χ^2 testing. Significant effects are marked with *. Main effects are significant at $\alpha = 0.05$ and post-hoc effects are significant at $\alpha = 0.016$.

As in Section 3.2.1, we find that the sv-PPA versus controls task has the highest number of distinguishing features. In some ways this is counter-intuitive, as we might expect these grammar-based features to be more relevant to the two nfv-PPA classification tasks. However, a closer look at the selected features reveals that the features capture some semantic informa-

Rule	controls (total)	sv-PPA (total)	controls (mean)	sv-PPA (mean)
VP → VBD SBAR	59	11	3.69	1.00
VP → VBD S	42	9	2.63	0.82
VP → AUX RB VP	29	40	1.81	3.64
VP → VBD NP PP	21	3	1.31	0.27
VP → VBG PP	15	28	0.94	2.55
VP → VBZ PP	14	1	0.88	0.09
VP → VBP	5	12	0.31	1.09
NP → PRP	690	543	43.13	49.36
NP → NNP	54	16	3.38	1.45
NP → DT NN	349	166	21.81	15.09
NP → DT NN NN	35	4	2.19	0.36
NP → PRP$ NN NN	11	0	0.69	0.00
NP → DT	33	59	2.06	5.36
NP → EX	9	23	0.56	2.09
PP → IN NP	322	251	20.13	22.82
PP → TO NP	111	35	6.94	3.18
PP → IN PP	1	8	0.06	0.73

(a) Control versus sv-PPA features.

Rule	controls (total)	nfv-PPA (total)	controls (mean)	nfv-PPA (mean)
VP → VBN PP	33	6	2.06	0.46
VP → AUX SBAR	9	16	0.56	1.23
NP → DT	33	52	2.06	4.00
NP → EX	9	17	0.56	1.31
NP → NNP NNP	2	9	0.13	0.69
NP → DT NP CC NP	1	6	0.06	0.46

(b) Control versus nfv-PPA features.

Rule	sv-PPA (total)	nfv-PPA (total)	sv-PPA (mean)	nfv-PPA (mean)
VP → AUX VP	90	49	8.18	3.77
VP → VBD SBAR	11	24	1.00	1.85
VP → VBD S	9	22	0.82	1.69
VP → VBG NP	17	3	1.55	0.23
NP → NNS	28	8	2.55	0.62
PP → TO NP	35	57	3.18	4.38

(c) sv-PPA versus nfv-PPA features.

Table 3.13: Features with a standardized residual greater than 2.5 in the χ^2 test (i.e. those features which contribute to the lack of fit between the distributions).

tion, as well. To help interpret the results in Table 3.13, we present some examples of these constructions from the data, and discuss their relation to the expected language impairments in each subtype.

Controls vs sv-PPA

The first two rules from Table 3.13a show that controls are more likely than sv-PPA patients to create a verb phrase from a past-tense verb (VBD) and an independent or dependent clause (S or SBAR, respectively). An example of this construction from the control data is: *Cinderella* ***realized it was midnight***. One question that arises is whether sv-PPA participants are less likely to attach clauses to verbs, or whether they are less likely to use past-tense verbs in general. If we consider *all* cases where an S-clause occurs after a verb in a verb phrase, we find a greater number in the control data (81 cases for controls, average: 5.06 per narrative; 42 cases for sv-PPA, average: 3.82 per narrative). For SBAR-clauses occurring after a verb in a verb phrase, we also find a greater number in the control data (154 cases for controls, ave: 9.63; 71 for sv-PPA, ave: 6.45). Considering instead the question of verb tenses, we find many more verb phrases containing past tense verb forms (VBD) in the control data (240 examples for controls, ave: 15.0; 94 for sv-PPA, ave: 8.54). So it would appear that sv-PPA participants are less likely to use past-tense verbs *and* less likely to attach clauses to verbs in general.

The next rule, VP → AUX RB VP, occurs more often in sv-PPA narratives than in control narratives. In some cases, this represents narrative content, such as *She* ***wasn't going*** (where *was* is the auxiliary and the negation is the adverb). However, in 60% of cases (24 out of 40, as labelled by hand), the construction is used in an expression of semantic difficulty, such as ***I don't know*** or ***I didn't see him***.

We then have three verb phrases involving prepositional phrases. VP → VBD NP PP and VP → VBZ PP are used more frequently by controls, while VP → VBG PP is used more frequently in the sv-PPA group. Meteyard et al. (2014) found that patients with sv-PPA produced significantly more verbs in *-ing* forms than controls, and the same effect could be driving the results here, particularly given our observation that sv-PPA patients are less likely to use *-ed* endings.

The last VP rule is VP → VBP, or verb phrases consisting of only a present-tense verb. While these constructions are not common in either group, they are more common in the sv-PPA group. Examining the data, one sv-PPA utterance using this construction is a narrative statement (*away they **go***), while the rest are discourse markers like *you **know*** and hedging phrases like ***I think*** or ***I guess***.

Turning to the noun phrases, we see that sv-PPA participants use more pronouns and fewer proper nouns per narrative than controls, as expected. They also use fewer noun phrases consisting of a determiner and a noun, and fewer noun phrases consisting of a determiner and two nouns. Examining the second case in more detail, we find that the two most common instantiations of NP → DT NN NN are references to ***the fairy godmother*** or ***the glass slipper(s)***. Only one of the sv-PPA participants uses the word *glass* to describe Cinderella's shoes (and as mentioned previously, none of the sv-PPA participants refer to Cinderella's footwear as *slippers*). Furthermore, none of the sv-PPA participants describe the fairy godmother as such, instead describing her as simply a *fairy*, or in some cases a *witch* or an *angel*. The rule NP → PRP$ NN NN is another example of this phenomenon.

In contrast, we find that sv-PPA participants are more likely to use noun phrases consisting of only a determiner. In most cases this corresponds to use of a demonstrative pronoun, e.g., *stuff like **that*** or *he get **this***. Meteyard et al. (2014) also found that sv-PPA participants used significantly more demonstrative pronouns than controls. Participants with sv-PPA are also more likely to use EX, the "existential there"; examples from the sv-PPA group include ***there****'s one type of small animals* and ***there*** *was a party coming on*.

Finally, we consider prepositional phrases. PP → TO NP is used more frequently in control narratives than sv-PPA narratives, while the other two PP constructions are more common (on average) in the sv-PPA groups. In total the control group uses 453 (ave: 28.31) prepositional phrases, and the sv-PPA group uses 317 (ave: 28.81). So sv-PPA and control participants use approximately the same number of prepositional phrases, on average. However, out of those totals, 25% of prepositional phrases produced by controls take the form of PP → TO NP, and only 11% of those produced by sv-PPA participants do. The difference between this rule and all the other rules with PP on the left-hand side is that the preposition is tagged with TO (indicating the word *to*) rather than IN (indicating a preposition or subordinating conjunction more generally). In particular, *to* is a preposition of movement: it expresses a movement from one place *to* another. Some examples from the control data include *they went **to Cinderella's house*** and *she went **to the ball***. Since all the other prepositions are grouped together under the category IN, we cannot say whether this represents a specific deficit for *to* prepositions in sv-PPA, or whether perhaps there is another preposition which is favoured over all others.

Controls versus nfv-PPA

For controls versus nfv-PPA, there are two verb phrase constructions in Table 3.13b. The first takes the form VP → VBN PP. Examination of the parse trees reveals that 6 of the 33 productions of this rule in the control group (18%) were due to parse errors where past tense verbs were mistakenly tagged as past participles – an easy mistake, since they take the same form for regular verbs. The correctly tagged instances correspond to phrases like *she was **relegated to doing all the menial tasks*** and *they were **turned into horses***. However, in the nfv-PPA group, only one instance of the rule accurately corresponds to use of the past participle. If we count up the total frequency of VBN occurrences in the data, we find 52 cases in the control data and only 14 in the nfv-PPA data. Given that the correct use of a past participle involves both an auxiliary verb and an inflected form, this may reflect an underlying agrammatism or syntactic simplification in the nfv-PPA group.

VP → AUX SBAR occurs more frequently in the nfv-PPA group. In 3/16 of the cases, this corresponds to a grammatical construction (e.g., *that**'s who she married***). In the majority of cases, however, it actually corresponds to a false start and subsequent repair, such as *she **is the she was a slave*** and *they **were it was all about them***.

The first two NP rules in Table 3.13b involve the use of demonstrative pronouns and "existential there", as discussed above in the sv-PPA versus control case. Again, these constructions are used more frequently in the patient data, suggesting that they are features of PPA speech in general.

The last two rules in the nfv-PPA versus controls case have very small frequency counts, so their generalizability is unclear, but in this corpus nfv-PPA participants use NP → NNP NNP and NP → DT NP CC NP more than controls. The first case corresponds to multiple attempts at the word *Cinderella*, or non-word paraphasias in general (when a word is not in the parser's vocabulary, it often resorts to tagging it as a proper noun). The second case occurs 6 times in the nfv-PPA group, and every time corresponds to a repeated determiner (e.g., ***the the footwear and the prince***). Note that if the determiner wasn't repeated, then matching rule would be simply NP → NP CC NP, which occurs 21 times in the nfv-PPA group (ave: 1.62) and 46 times in the control group (ave: 2.88).

Differentiating PPA subtypes

On average, the construction VP → AUX VP occurs over twice as often per-narrative in the sv-PPA group as in the nfv-PPA group. Many of the examples correspond to the past progressive tense (*she* ***was working***, *she* ***was running out***). This is consistent with (1) an increased reliance on *-ing* forms in sv-PPA, and (2) decreased production of function words, such as auxiliaries, in nfv-PPA.

The next two VP rules (VP → VBD SBAR and VP → VBD S) correspond to the first two rules in Table 3.13a, and have been discussed at length above. Again, we observe that these constructions are used less frequently in the sv-PPA group relative to the nfv-PPA group. This suggests that the effect is due to the notable lack of such constructions in sv-PPA rather than an increase in nfv-PPA (in fact, nfv-PPA production of these constructions is also reduced relative to controls, but to a lesser degree).

The final VP construction takes the form VP → VBG NP, and is more frequent in the sv-

PPA group. Unfortunately, it appears that 6/17 of those cases are due to a parser error, where the filled pause *um* is mis-tagged as VBG.[4] The remaining 11 cases are either past or present progressive (e.g., *she was* ***making clothing***, *they're* ***doing something***), in line with previous results from this section.

Looking at NP → NNS, it appears that participants in the sv-PPA group use more plural nouns without determiners. Figure 3.9 shows that sv-PPA participants use this construction more than controls, and nfv-PPA participants use this construction far less than controls. If we include all constructions with NNS (i.e. including NP → DT NNS and NP → PRP$ NNS), the trend still holds: controls use an average of 4.25 plural nouns per narrative, sv-PPA participants use 5.6, and nfv-PPA participants use only 2.8. This deficit for plural nouns in nfv-PPA may be related to an impairment in grammatical morphology.

Finally, the last entry in Table 3.13c represents the paucity of PP → TO NP structures in the sv-PPA data, as discussed above. Again, the frequency of this structure is also reduced in the nfv-PPA group relative to controls, but not to the extent that is seen in sv-PPA. Therefore, this should be interpreted as a feature whose reduction distinguishes sv-PPA in general, rather than whose increase distinguishes nfv-PPA.

3.4.3 Phrase-level statistics

We also consider a set of phrase-level statistics. These are a subset of the features described by Chae and Nenkova (2009), and are related to the incidence of different phrase types. We again consider three different phrase types: noun phrases, verb phrases, and prepositional phrases. These features are defined as follows:

- **Phrase type proportion** Total number of words in each phrase type (including embedded phrases), divided by total number of words in the narrative.

[4] The reason for this is unclear, since present participles end with "ing" in English, but is presumably related to the parser trying to manipulate ungrammatical text to fit a more probable construction.

- **Average phrase length** Total number of words in each phrase type, divided by number of phrases of that type.

- **Phrase type rate** Number of phrases of a given type, divided by total number of words in the narrative.

Because we are interested in the grammaticality of the entire narrative, we normalize by narrative length (rather than sentence length, as in Chae and Nenkova's study). These features are real-valued.

When we compare the group means on each of these measures, a two-tailed t-test reveals an uncorrected significant difference on sv-PPA versus controls for VP proportion (sv-PPA mean 2.06, control mean 2.89, $p = 0.03$), average VP length (sv-PPA mean 9.67, control mean 13.09, $p = 0.04$, and average NP length (sv-PPA mean 2.99, control mean 3.83, $p = 0.03$). However, none of these differences are significant if we correct for multiple comparisons.

3.4.4 Discussion

While detailed analysis of the sort presented above can help contextualize individual CFG features, it can be easy to lose sight of the overall picture. Here, we will summarize some of the key findings from this section at a higher level.

Syntactic productions can reflect semantic impairment

There were a number of different features which distinguished the sv-PPA group from controls. Many of the phenomena observed here seem to relate more clearly to a semantic deficit than a syntactic one, including an increase in the number of personal pronouns and demonstrative pronouns, a decrease in proper nouns and noun phrases corresponding specifically to story elements like *the fairy godmother* and *the glass slippers*, and an increase in structures corresponding to extra-narrative comments on word-finding difficulties, like *I don't know*. In some sense, these production rules may be seen as an extension of our previous approach of counting individual parts-of-speech, except that we now consider syntactic labels higher in the hierarchy than individual POS tags. Since it has already been established that the distribution of POS tags changes in the presence of semantic impairment (e.g., increasing verbs, decreasing nouns), it is unsurprising that the distributions of higher-order syntactic structures are also affected.

Potential syntactic changes in sv-PPA

There were, however, some constructions which appeared to point to syntactic, rather than semantic, changes in sv-PPA. Specifically, sv-PPA participants used *-ing* verbs in progressive tenses more frequently than controls or nfv-PPA participants, and showed a corresponding deficit in other verb forms. They also showed a reduction in prepositional phrases consisting of the preposition *to* followed by a noun phrase. Previous work has also found an increase in *-ing* verbs in sv-PPA (Meteyard et al., 2014), although the potential relevance of the second result will require further investigation.

While the extent to which grammar is compromised in sv-PPA is still an open research question, other researchers have theorized that differences in semantics will result in differences in the syntactic structures which are selected in narrative speech. Meteyard and Patterson (2009) suggest that as the disease progresses, sv-PPA patients may rely heavily on highly frequent structures and routinized phrases, paralleling their reliance on highly frequent lexical items. They also found evidence of syntactic errors, concluding that “Whilst there was no evidence of gross syntactic violations, there was definitely an increased rate of errors on free and bound closed class items, and syntactic anomalies occurred when lexical retrieval went awry.” Meteyard et al. (2014) found that their sv-PPA patients showed a reduced range of complex structures, and a different distribution of constructions relative to controls, much as we did here.

Relationship between speaker errors and parser errors

In the nfv-PPA case, many of the relevant features corresponded to dysfluencies produced by the nfv-PPA speakers (e.g., repetitions, false starts). Modern statistical parsers take a rather descriptionist view of grammar: they generally will not fail on an ungrammatical sentence, they will simply apply the rules of the grammar as best they can to produce the most probable parse of the sentence. In some cases, this results in the use of low-frequency rules, as we saw in the case of NP → DT NP CC NP. These low-frequency rules may be used to detect the occurrence of dysfluencies in the samples.

The other side of this issue is that the parser often makes mistakes when the input does not conform to the expected grammar. For example, only 1/6 cases of the production VP → VBN PP in nfv-PPA actually contained a past participle verb. A future direction for analysis of this type could be to associate a confidence score to each structure, as in Lin and Weng (2008), to help determine the trustworthiness of the parser output.

Distinguishing between subtypes

One motivation for introducing these features was to find additional features that distinguish between the two subtypes of PPA. Interestingly, while some of the features which distinguished nfv-PPA from controls did point to the occurrence of dysfluencies in the nfv-PPA group, these features did not distinguish between sv-PPA and nfv-PPA. Similarly to Section 3.2, most of the relevant features seem to be related to deficits in sv-PPA rather than nfv-PPA, although some of the features did hint at difficulties with grammatical morphology in nfv-PPA.

It has been suggested that narrative speech is not the most sensitive measure of syntactic impairment, since people in general (healthy and impaired) tend to naturally use simpler structures in such situations (Benedet et al., 2006; Meteyard and Patterson, 2009; Wilson et al., 2010). Despite identifying a number of differing features between the subtypes, this analysis did not conclusively point to an obvious syntactic deficit in nfv-PPA relative to sv-PPA. However, the practical utility of these features has yet to be tested in a classification task (see

Section 3.5, below).

Conclusion

The analysis presented in this section is limited in a number of ways: we have only a small sample of data, it involved an arbitrary cut-off of the 100 most frequent productions per group and then a semi-arbitrary cut-off of 2.5 for the magnitude of the Pearson residual, we limited our analysis to NPs, VPs, and PPs, and some of the results seem to correspond specifically to the task of Cinderella story-telling with unclear generalizability to other forms of speech production. However, we can conclude that there are differences in the CFG productions produced by the patient groups, and that in most cases they appear to be associated with expected changes in language in PPA. In the next section, we will explore the utility of these features in the classification framework.

3.5 A comparison of feature sets for PPA classification[5]

At this point, we have a large number of features from both the transcripts and the associated audio files. To conclude this chapter, we conduct a thorough analysis of which *types* of features are most useful to the classification tasks. This is a matter of both theoretical and practical interest. To learn more about PPA, we would like to know which aspects of language are most affected by the disease: The number of unique words used? The syntactic classes of the words? The structure of the sentences? The fluency of speech? and so on. However, we would also like to know if there are some features which simply are not worth measuring, either because they measure characteristics of speech or language which are not affected in PPA, or because they are too noisy or inconsistent to be of value.

We perform two experiments. We first evaluate the individual feature sets on their classi-

[5] The material presented in this section was previously published in: Kathleen C. Fraser, Graeme Hirst, Naida L. Graham, Jed A. Meltzer, Sandra E. Black, and Elizabeth Rochon (2014). Comparison of different feature sets for identification of variants in progressive aphasia. *Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, Baltimore, Maryland.

fication accuracy, and then perform an ablation study to determine the optimal combination of feature sets.

3.5.1 Feature sets

The features in this section have been previously introduced; here we simply describe the grouping of the features into different feature sets. Table 3.14 presents the seven feature sets, the abbreviation for the sets, and a description of the features assigned to each set. In many cases the groupings are self-explanatory; however, a few features were somewhat ambiguous. We assigned word length to the "Complexity" set, with the rationale that it measures word-level complexity. The total number of words was assigned to the "Fluency" set. The "Vocabulary richness" set includes a variation on type:token ratio (TTR) called *moving average type:token ratio* (MATTR), introduced by Covington and McFall (2010) as an alternative to TTR which is not dependent on sample length.

Acoustic (A) Features derived directly from the audio file, e.g., mean pause duration, fundamental frequency variance.

Complexity (C) Features capturing some notion of syntactic complexity, e.g., mean Yngve depth, mean length of utterance.

CFG production rules (CFG) Features derived from frequency counts of CFG constituents and phrase-level features, e.g., NP → DT NN, VP proportion.

Fluency (F) Features measuring aspects of fluency from the transcripts, e.g., number of filled pauses, number of NID words.

Psycholinguistic (P) Features relating to the psycholinguistic properties of words, e.g., frequency, familiarity.

Part-of-speech (POS) Features measuring the distributions of different POS categories, e.g., noun:verb ratio, number of adverbs.

Vocabulary richness (VR) Features measuring the lexical diversity of the sample, e.g., type:token ratio, Brunét's index.

Table 3.14: Feature set names, abbreviations (in parentheses), and descriptions.

3.5.2 Experiments

We report the results of two experiments exploring the discriminative power of the different features. We first compare the classification accuracies using each individual feature set. We then perform an ablation study to determine which combination of feature sets leads to the highest classification accuracy. Unlike in previous sections, we do not perform feature selection; rather we include or exclude entire sets of features at a time.

Accuracies using individual feature sets

The accuracies which result from using each feature set individually are given in Table 3.15. The highest accuracy across the three tasks is achieved in distinguishing sv-PPA participants from controls. An accuracy of .963 can be achieved using all the features together, or using the psycholinguistic or POS features alone. This is consistent with the semantic impairments that are observed in sv-PPA. The measures of vocabulary richness do not distinguish between the sv-PPA and control groups, suggesting it is the words themselves, and not the number of different words being used, that is important.

In the case of nfv-PPA participants vs. controls, we find that the highest accuracy of .931 is achieved using all the features, and the second highest (.862) by using only acoustic features. This suggests that the results of the PPA vs. controls experiment in Section 3.3 may have been driven primarily by the nfv-PPA group (note that the accuracy using acoustic features to detect sv-PPA is lower, at .778). The third best accuracy is achieved using the fluency features. Both acoustic and fluency features could potentially detect the hesitant, halting speech that is characteristic of nfv-PPA. Once again, the complexity and CFG features are not particularly sensitive to this classification task.

Finally, the best accuracy for sv-PPA vs. nfv-PPA is lower than in the previous two cases, and is achieved using only CFG features. This indicates that there are some grammatical constructions which occur with different frequencies in the two groups, as discussed in Section 3.4. These differences do not appear to be captured by the complexity features, which explains why

Feature set	sv-PPA vs. controls	nfv-PPA vs. controls	sv-PPA vs. nfv-PPA
All	**.963**	**.931**	.708
Acoustic	.778	.862	.167
Psycholinguistic	**.963**	.724	.708
POS	**.963**	.690	.375
Complexity	.852	.621	.667
Fluency	.667	.828	.500
Vocab. richness	.481	.586	.583
CFG	.630	.690	**.792**

Table 3.15: Classification accuracies for each feature set individually using a SVM classifier. Bold indicates the highest accuracy for each task.

we did not find syntactic differences between the sv-PPA and nfv-PPA groups in Section 3.2. Interestingly, the results using CFG features are actually higher than the results using all features. This demonstrates that classifier performance can be adversely affected by the presence of irrelevant features, especially in small data sets.

Combining feature sets

In the previous subsection we examined the feature sets individually; however, one type of feature may complement the information contained in another feature set, or it may contain redundant information. To examine the interactions between the feature sets, we perform an ablation study. Starting with all the features, we remove each feature set one at a time and measure the accuracy of the classifier. The feature set whose removal causes the smallest decrease in accuracy is then removed permanently from the experiment, the reasoning being that the most important feature sets will cause the greatest decrease in accuracy when removed (Bethard, 2007). In some cases, we observe that the classification accuracy actually *increases* when a set is removed, which suggests that those features are not relevant to the classification (at least in combination with the other sets). In the case of a tie, we remove the feature set whose individual classification accuracy on that task is lowest. The procedure is then repeated on the remaining feature sets, continuing until only one set remains.

The results for sv-PPA vs. controls are given in Table 3.16a. The best result, 1.00, is

Removed	Remaining Features	Accuracy
	A+P+POS+C+F+VR+CFG	.963
F	A+P+POS+C+VR+CFG	.963
A	P+POS+C+VR+CFG	1.00
VR	P+POS+C+CFG	.926
CFG	P+POS+C	.926
C	**P+POS**	**1.00**
POS	P	.963

(a) sv-PPA vs. controls.

Removed	Remaining Features	Accuracy
	A+P+POS+C+F+VR+CFG	.931
VR	A+P+POS+C+F+CFG	.931
C	A+P+POS+F+CFG	.931
POS	A+P+F+CFG	.931
CFG	A+P+F	.966
F	**A+P**	**.966**
P	A	.862

(b) nfv-PPA vs. controls.

Removed	Remaining Features	Accuracy
	A+P+POS+C+F+VR+CFG	.708
POS	A+P+C+F+VR+CFG	.750
VR	A+P+C+F+CFG	.833
F	A+P+C+CFG	.833
A	P+C+CFG	.792
C	**P+CFG**	**.917**
P	CFG	.792

(c) sv-PPA vs. nfv-PPA.

Table 3.16: Ablation study results. A=acoustic, P=psycholinguistic, POS=part-of-speech, C=complexity, F=fluency, VR=vocabulary richness, CFG=CFG production rule features. Bold indicates the highest accuracy with the fewest feature sets.

achieved by combining the psycholinguistic and POS features. This is unsurprising, since each of those feature sets perform well individually. Curiously, the same result can also be achieved by also including the complexity, vocabulary richness, and CFG features, but not in the intermediate stages between those two optimal sets. We attribute this to the interactions between features and the small data set.

For nfv-PPA vs. controls, shown in Table 3.16b, the best result of .966 is achieved using a combination of acoustic and psycholinguistic features. In this case the removal of the fluency features, which gave the second highest individual accuracy, does not make a difference to the accuracy. This suggests that the fluency features contain similar information to one of the remaining sets, presumably the acoustic set.

In the case of sv-PPA vs. nfv-PPA, we again see that the best accuracy can be achieved by combining two feature sets, as shown in Table 3.16c. Using psycholinguistic and CFG features, we can achieve an accuracy of .917, a substantial improvement over the best accuracy for this task in Table 3.15. In fact, in all three cases we see that using a carefully selected combination of feature sets can result in better accuracy than using all the feature sets together or using any one set individually.

3.5.3 Discussion

In the first experiment, we found that the single best feature set for distinguishing between sv-PPA and controls was either psycholinguistic features (e.g., frequency, familiarity) or part-of-speech features, the best feature set for distinguishing between nfv-PPA and controls was acoustic features, with fluency features as a close second, and the best feature set for distinguishing between sv-PPA and nfv-PPA was the CFG feature set. In the second experiment we found that by combining feature sets in a systematic manner, we could achieve better classification results than either choosing any one feature set, or combining all the feature sets together. Notably, our best result for classifying sv-PPA vs. nfv-PPA was achieved using a combination of CFG and psycholinguistic features. The accuracy in that case was .917, a considerable increase over the best accuracy (.792) on that task in Section 3.2. We conclude that while it may be tempting to calculate as many features as possible and use them all in a classifier, better results can be achieved by choosing a small, relevant subset of features.

3.6 Summary

In this chapter, we showed that we can extract a large quantity of relevant information from a short sample of patient speech. In the first section, we extracted a variety of syntactic, psycholinguistic, and fluency-based features from the manual transcriptions of the *Cinderella* stories. Using these features, we could achieve high classification accuracies for distinguishing between PPA participants and controls, but lower accuracy for distinguishing between the semantic and nonfluent/agrammatic variants. Surprisingly, the syntactic features were not significantly different in the nfv-PPA group relative to the sv-PPA group. A partial least squares analysis showed good separation between all three groups in two dimensions. Importantly, the features which were identified as being relevant to the distinction between groups were consistent with previous literature in the field.

We then considered adding acoustic features to the classification pipeline. One benefit of acoustic features is that they do not require a transcription of the stories; however, they proved to be useful only in distinguishing between PPA patients and controls, and not between the two subtypes. On the task of detecting PPA, important features included long and short pause counts, mean duration of pauses, and phonation rate, which point to the word-finding difficulties and dysfluencies that can be seen in PPA.

We then investigated a new set of features to analyze the syntactic changes in PPA in more detail. By counting the frequency of occurrence of different CFG productions, we could identify the syntactic structures that were used more (or less) often in each group. Although formulated as syntactic features, we found that some production rules were associated with semantic phenomena. We also linked certain low-frequency rules to dysfluencies such as repetitions or false starts.

Finally, we compared the performance of an SVM classifier when trained on different sets of the features presented in the chapter. We first compared the accuracies that could be achieved using each set of features individually, and found that the results varied depending on the classification task. Psycholinguistic, POS, acoustic, fluency, and CFG features were all identified as important, indicating that different features types capture different linguistic impairments. An ablation study revealed that a combination of psycholinguistic and POS features was most successful at distinguishing sv-PPA from controls, a combination of psycholinguistic and acoustic features was most successful at distinguishing nfv-PPA from controls, and a combination of psycholinguistic and CFG features was most successful at distinguishing between the two subtypes of PPA.

The automated extraction of features from a transcript is incredibly time-saving relative to manual annotation. Features as specific as the CFG features, which require parsing every utterance in the transcript, are practically infeasible to extract manually. However, relying on manual transcription and utterance segmentation is a major obstacle to creating a fully-automated system of analysis. In the next chapter, we will consider the effects of incorporating automatic speech recognition and automated boundary segmentation.

Chapter 4

Classification of PPA from real and simulated ASR transcripts

It is clear that automatic speech recognition (ASR) will be necessary to create a fully automated, end-to-end analysis pipeline. However, as discussed in Section 2.6, speech recognition systems do not tend to work well for older adults in general, and people with dementia in particular. In this chapter, we present the results of an experiment using off-the-shelf ASR software to recognize the speech samples in our data set, followed by an analysis of the effect of the noisy recognition on the features, and on the subsequent classification accuracy.

In the second half of this chapter, we consider the related problem of automated sentence segmentation. Since the output of an ASR system is typically a stream of text, an automatic method for segmenting that text into meaningful syntactic units is required. In this study, we train a classifier to label each interword boundary as either a sentence boundary or not, and compare the results against the manual sentence boundary annotations. We also compare the accuracy with the results we obtain on a corpus of news data. Once again, we consider the effect of the noisy segmentation on relevant features, and the performance of classifiers trained on such features.

4.1 A preliminary exploration of automatic speech recognition for PPA classification[1]

Fully automated analysis of narrative speech will require automatic speech recognition (ASR) in order to extract lexical and syntactic features from acoustic signals. Despite major improvements in ASR technology over the past few decades, accuracy for unrestricted (i.e., 'dictation-style') speech remains decidedly imperfect, and accuracy for older and impaired speakers is often much worse. In order to estimate how effective a classifier of PPA and its subtypes might be given textual transcripts derived from ASR, a wide range of potential system performances must be considered, to account for real-world variation. This research approximates various levels of ASR performance by randomly corrupting human transcripts according to pre-defined levels of error and compares these results against actual output from a leading commercial dictation system. Error levels are quantified by word-error rate (WER), which is the number of erroneous insertions, deletions, and substitutions of words in an ASR transcript, divided by the total number of words in a reference transcript[2]. Simulated ASR errors have been used in various contexts, such as training dialogue systems (Schatzmann et al., 2007) and for testing the safety of dictation systems for use in automobiles (Labský et al., 2012).

4.1.1 Features

Two sources of information are available for each participant individually, namely text transcripts and acoustic samples. From these, we derive lexical/syntactic features from the transcripts (features 1 and 27–58 in Table 3.2) and acoustic features from the audio samples (see Table 3.8), giving a total of 54 available features.

[1] The material presented in this subsection was previously published in: Kathleen Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon (2013). Automatic speech recognition in the diagnosis of primary progressive aphasia. *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pp. 47–54

[2] If the number of insertions is large, it can overwhelm the total number of words in the reference transcript, thereby allowing for WERs above 100%.

In Chapter 3, when using manual transcripts, we examined measures which can be derived from parse trees, such as Yngve depth, or the number and length of different syntactic constructions. However, such parse trees will depend on the location of the sentence boundaries in the transcript, the placement of which can be a difficult task for ASR systems (Liu et al., 2006). Indeed, the Nuance system used here does not place punctuation except by explicit command. For the purposes of this preliminary study, we avoid using features which depend on accurate sentence boundaries.

4.1.2 ASR and simulated errors

We use two methods to produce errorful textual transcripts. The first method represents the current leader in commercial dictation software, Nuance Dragon NaturallySpeaking Premium; here, audio files are transcribed by Nuance's desktop dictation software. The second method corrupts human-produced transcripts according to pre-defined levels of WER; this method allows for an indirect approximation of the performance given a wide range of potential alternative ASR systems.

The Nuance Dragon NaturallySpeaking 12.5 Premium for 64-bit Windows dictation system (hereafter, 'Nuance') is based on traditional hidden Markov modeling of acoustics and, historically, on trigram language modeling (Francois, 2008). This system is initialized with the default 'older voice' model suitable for individuals 65 years of age and older. The default vocabulary consists of 150,478 words, plus additional control phrases for use during normal desktop dictation (e.g., "*new paragraph*", "*end of sentence*"); this feature cannot be deactivated. The core vocabulary, however, can be changed. In order to get a more restricted vocabulary, all words used in the Cinderella data set plus all words used in a selection of 9 stories about Cinderella from Project Gutenberg (totalling 22,168 word tokens) were combined to form a reduced vocabulary of 2633 word types. Restricted vocabularies, by their nature, have higher random baselines and fewer phonemically confusable word pairs, usually resulting in proportionally higher accuracies in ASR.

For the simulated ASR transcripts, each word in the manual transcript is modified with a probability equal to the desired WER. A word w can be modified in one of three ways:

- Substitution – w is replaced with a new word w_S.
- Insertion – w is followed by a new word w_I.
- Deletion – w is removed.

In the case of insertion, the word to be inserted is chosen according to the bigram distribution of the language model obtained on the Gigaword corpus (Graff and Cieri, 2003). If w is not found in the vocabulary, then w_I is chosen according to the unigram distribution of the language model. In the case of substitution, the new word is randomly chosen from a ranked list of words with minimal phonemic edit distance from the given word, as computed by the Levenshtein algorithm.

Once it has been determined that a word will be modified, it is assigned one of the above modifications according to a pre-defined distribution. Different ASR systems may tend towards different distributions of insertion errors (IE), substitution errors (SE), and deletion errors (DE). We create data noise according to three distributions, each of which favours one type of error over the others: [60% IE, 20% SE, 20% DE], [20% IE, 60% SE, 20% DE], and [20% IE, 20% SE, 60% DE]. We then also adjust these proportions according to proportions observed in Nuance output, as described in Section 4.1.4.

4.1.3 Classification

We use leave-one-out cross-validation to test our diagnostic classifiers. For each fold, one transcript is removed as test data. We then apply a simple feature selection algorithm to the remaining transcripts: we calculate a Welch's t-test for each feature individually and determine the significance of the difference between the groups on that feature. We then rank each feature by increasing p-value, and include as input to the classifier only the top ten most significant features in the list. For each fold, different training data is used and therefore different features may be prioritized in this manner.[3]

Once the features have been selected, we train three types of classifier: naïve Bayes (NB), support vector machine with sequential minimal optimization (SVM), and random forests (RF). The classifiers are then tested on the held-out transcript. This procedure is repeated for every transcript in the data set, and the average accuracy is computed.

We consider two classification tasks, PPA vs. controls and sv-PPA vs. nfv-PPA, since these binary classification tasks can be cascaded. For each task, there are two possible feature sets: text features only, or a combination of text and acoustic features. There are also two possible training sets for each task: i) the classifiers can be trained on the human-transcribed data and tested on the ASR data (for example, if researchers have access to a corpus of manual transcriptions for training purposes), and ii) the classifiers are both trained and tested on the noisy ASR (or simulated ASR) data. We test our classifiers on each combination of these variables.

4.1.4 Results

Recognizing PPA speech

Table 4.1 shows the WER of the Nuance system across populations and vocabularies. Somewhat surprisingly, using the reduced vocabulary reduces accuracy considerably, despite all words in the test set being present in the vocabulary. A possible explanation may be found in the distribution of error types across the uses of both vocabularies, which is shown in table 4.2. In particular, Nuance makes significantly more deletion errors when using the reduced vocabulary, which may be attributed to a lower confidence associated with its word sequence hypotheses, which in turn may be attributed to a language model that is not adapted to non-default vocabularies. A general language model may assign a high lexical probability to a series of words that are phonemically similar to an utterance but if those words are not in the

[3] In the classification experiments in Chapter 3, feature selection occurred outside the cross-validation framework. However, a more accurate estimate of the system's performance on unseen data is obtained by performing the feature selection within the cross-validation.

	Default Vocabulary	Reduced Vocabulary
sv-PPA	73.1	98.1
nfv-PPA	67.7	97.3
Control	64.0	97.1
All	67.5	97.5

Table 4.1: Mean word error rates for the Nuance systems on each of the participant groups.

	Default Vocabulary	Reduced Vocabulary
Insertion errors	0.00602	0.00008
Substitution errors	0.39999	0.11186
Deletion errors	0.59398	0.88804

Table 4.2: Distribution of error types for the Nuance systems.

reduced vocabulary, a more domain-specific sequence of words may be assigned a low lexical probability and therefore a low confidence. When confidence in a hypothesis is below some threshold, that hypothesis may not be returned, resulting in an increase in deletion errors. Not having access to these internals of the Nuance engine prohibits modification at this level.

Another point to highlight is that, given Nuance's default vocabulary, there is no significant difference between the WER obtained with the control and nfv-PPA groups ($p = 0.54$), nor with the control and sv-PPA groups ($p = 0.16$).

Features and feature selection

Despite the high WER, some text features are still significant in the Nuance data. Table 4.3 shows the text features that were significant ($p < 0.05$) when comparing PPA and controls using the two Nuance models. Since the feature set changes with each fold in the cross-validation, the p-value is an average across folds. The means for the two groups are also shown to indicate the direction of the difference. Using the default vocabulary, there are three significant text features: verb imageability, noun frequency, and noun familiarity. These three features are all significant in the manually-transcribed data as well, and with the same direction. For the system trained on the reduced vocabulary, there are five significant text features, as indicated, only one of which (noun imageability) is not significant in the manual transcripts. All five features

	p-value	PPA mean	Control mean
Nuance default vocabulary			
verb imageability	0.0006	401	354
noun frequency	0.002	3.51	3.26
noun familiarity	0.04	575	558
Nuance reduced vocabulary			
average word length	0.003	5.44	6.21
noun frequency	0.006	3.13	2.77
noun imageability	0.01	487	554
noun familiarity	0.02	558	531
frequency	0.04	3.60	3.20

Table 4.3: Significant text features ($p < 0.05$) for PPA versus controls using the Nuance system with default and reduced vocabularies.

	p-value	sv-PPA mean	nfv-PPA mean
Nuance default vocabulary			
noun familiarity	0.002	596	560
familiarity	0.002	594	568
Nuance reduced vocabulary			
None	N/A	N/A	N/A

Table 4.4: Significant text features ($p < 0.05$) for sv-PPA versus nfv-PPA using the Nuance system with default and reduced vocabularies.

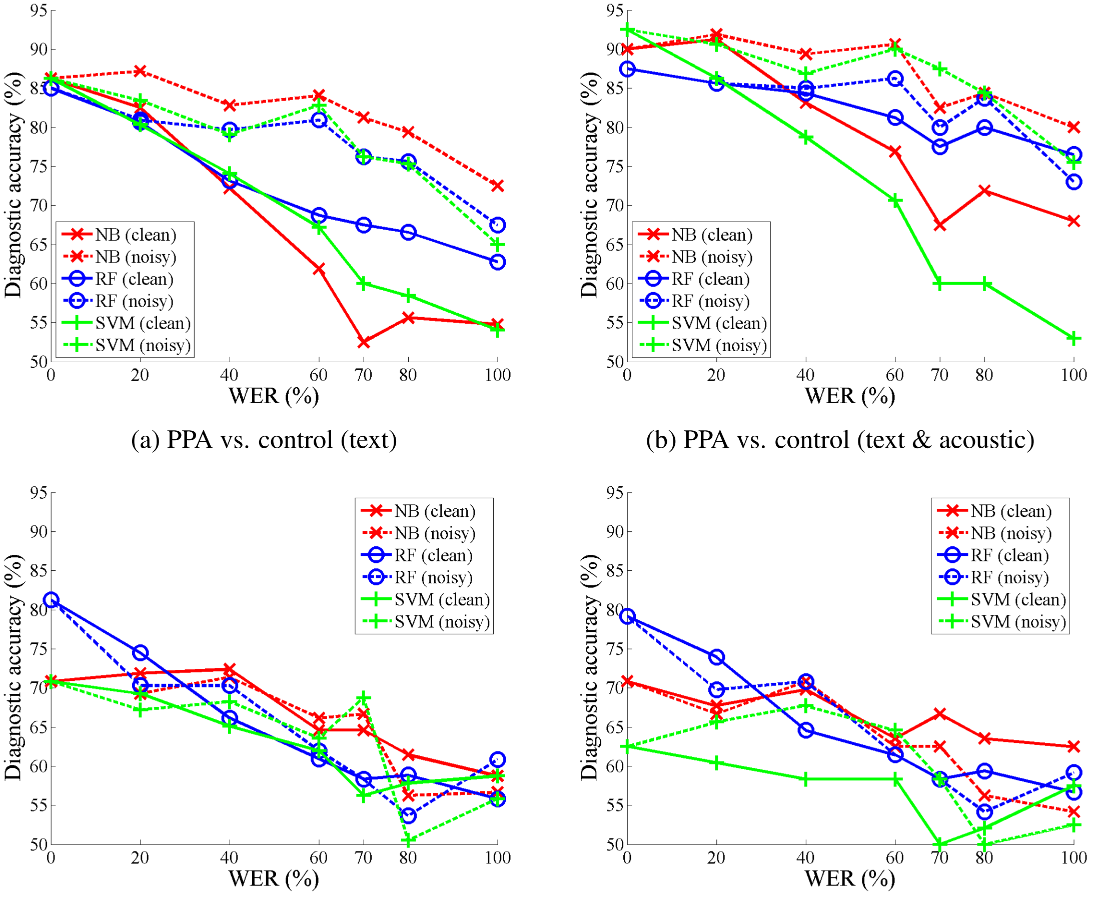
show differences in the same direction. Table 4.4 shows that only noun familiarity and overall familiarity are significant in the sv-PPA vs. nfv-PPA case using the default vocabulary system, as they are in the manually transcribed data, with the difference in the same direction. There are no significant text features using the reduced vocabulary system.

Diagnosing PPA and its subtypes

We evaluate the accuracy of diagnosing PPA and its subtypes based on the selected features across the three classification methods using the simulated ASR method. In practice, classification models might be trained on data that have been manually transcribed by human transcribers. However, as the amount of data increases, this becomes less practical and it may become necessary to train these models from transcripts that were automatically generated from ASR. We replicate our experiments once on data that have been manually transcribed and once on the same data, but with transcripts corrupted by synthetic word errors. Classifiers trained on human-produced transcripts had an average accuracy of 65.71% ($\sigma = 12.42$) and those trained on 'noisy' transcripts had an average accuracy of 70.72% ($\sigma = 13.89$), which is significant at $p < 0.00001$. These differences can be observed in Figure 4.1. Interestingly, the classifiers trained with 'noisy' transcripts outperform those trained with 'clean' transcripts fairly consistently in the PPA vs. control task, but this is far less pronounced (and to some extent reversed) in the sv-PPA vs. nfv-PPA task.

This trend is also apparent when the classifiers are tested using the Nuance transcripts. Figure 4.2 shows the classification accuracies for each classifier on each diagnostic task using the data generated using the default and reduced vocabularies. When classifying PPA versus controls, training on the 'noisy' Nuance data always leads to equal or greater accuracies than training on the 'clean' (human-transcribed) data. For sv-PPA versus nfv-PPA, the results are mixed, although the results from the reduced vocabulary suggest the opposite trend.

We compare the diagnostic accuracies across all classifiers given transcripts from Nuance using the reduced vocabulary with the accuracies of the synthetic WER method using the nearest WER (100%) and the associated error type distribution (i.e., 10% substitutions, 90% deletions, over all errors). We find no significant difference between results obtained with Nuance data and those obtained with the synthetic method ($p = 0.27$). We repeat this analysis with the default Nuance vocabulary and its equivalent synthetic WER (70%) and distribution (i.e., 40% substitution, 60% deletion) and again find no significant difference ($p = 0.15$). While not conclusive, the fact that there is no apparent difference in diagnosis when using the Nuance ASR and the synthetic method supports the use of the latter in these experiments.



(a) PPA vs. control (text)

(b) PPA vs. control (text & acoustic)

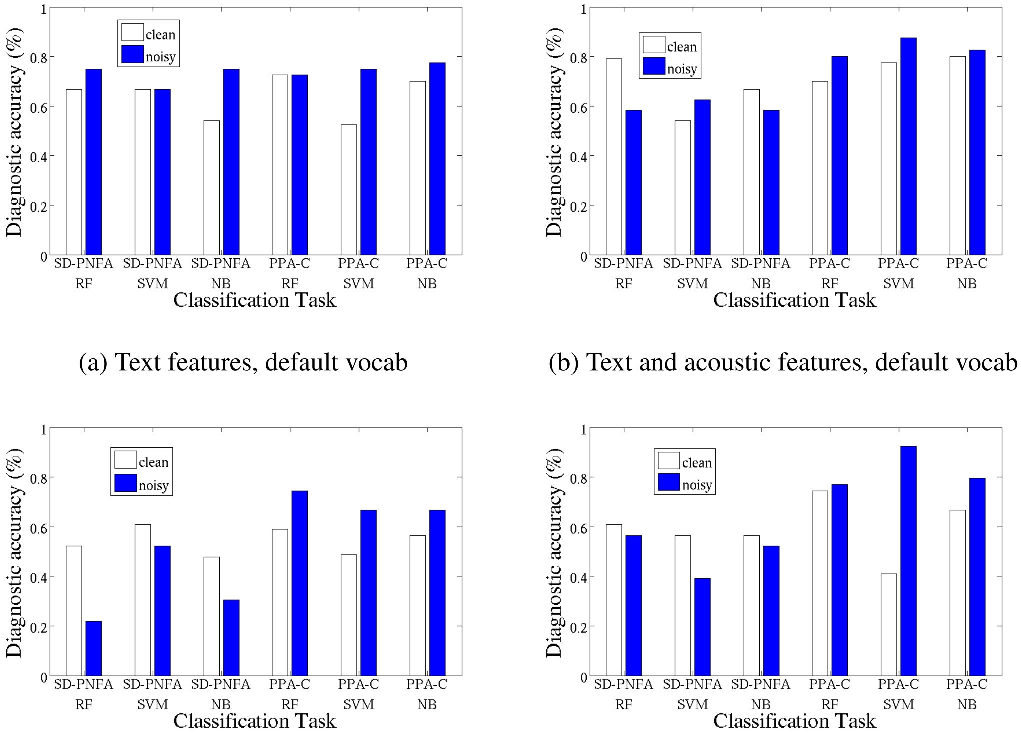
(c) sv-PPA vs. nfv-PPA (text)

(d) sv-PPA vs. nfv-PPA (text & acoustic)

Figure 4.1: Accuracy in diagnosing the indicated classes given features derived from potentially error-full textual transcriptions alone and in combination with features derived directly from the acoustics. Lines marked with x's, circles, and pluses indicate the use of the naïve Bayes, random forest, and support vector machine classifiers. Solid lines indicate those trained with human-transcribed data and dashed lines indicate those trained with corrupted data.

4.1.5 Discussion

This study represents an initial step towards incorporating ASR into a system for automated analysis of speech from people with possible dementia. One main result of this research is that classification of PPA can remain relatively accurate, even at very high levels of WER, by selecting appropriate features from the data at training time. Acoustic features are valuable, as they remain constant as the WER increases. However, our data suggest that some features from the text can still be informative, even when the transcripts are very noisy.



(a) Text features, default vocab

(b) Text and acoustic features, default vocab

(c) Text features, reduced vocab

(d) Text and acoustic features, reduced vocab

Figure 4.2: Classification accuracies for PPA versus controls and sv-PPA versus nfv-PPA using transcripts from the Nuance system with the default and reduced vocabularies. Empty bars indicate the accuracy achieved when training on the clean, human-transcribed data, while filled bars indicate the accuracy when training on the noisy ASR data.

One important direction for future work is to improve ASR for clinical populations. Here, we detect a trend towards worse ASR performance on PPA speech relative to speech from the general elderly population. This finding is in line with previous studies using ASR for participants with dementia (Peintner et al., 2008; Lehr et al., 2012), as discussed in Section 2.6. Our WER is especially poor for individuals with sv-PPA, suggesting that while more appropriate acoustic models built for older-adult voices will be important, a focus on improving language modeling may be more fruitful if the speaker has semantic impairments.

In this study we did not take into account any syntactic features, although agrammatism and/or syntactic simplification are characteristic of nfv-PPA. Including such information will require first applying a sentence boundary detection algorithm to the ASR transcripts, and then extracting traditional syntactic complexity measures, such as features 2–27 in Table 3.2. In the next section, we consider the problem of applying sentence segmentation algorithms to impaired speech.

4.2 Automatic sentence boundary detection in aphasic speech[4]

It is clear that automatic speech recognition (ASR) is required to build a fully automated analysis pipeline. However, as we discovered in the previous section, using ASR leads to another issue: the raw output from an ASR system is generally a stream of words, as shown in Figure 4.3. With some effort, it can be transformed into a format which is more readable by both humans and machines. Many algorithms exist for the segmentation of the raw text stream into sentences, but there has been no previous work on how those algorithms might be applied to impaired speech.

This problem must be addressed for two reasons: first, sentence boundaries are important when analyzing the syntactic complexity of speech, which can be a strong indicator of potential impairment. Many measures of syntactic complexity are based on properties of the syntactic parse tree (e.g., Yngve depth, tree height), which first require the demarcation of individual sentences. Even very basic measures of syntactic complexity, such as the mean length of sentence, require this information. Secondly, there are many reasons to believe that existing algorithms might not perform well on impaired speech, since assumptions about normal speech do not hold true in the impaired case. For example, in normal speech, pausing is often used to indicate a boundary between syntactic units, whereas in some types of dementia or aphasia a pause may indicate word-finding difficulty instead. Other indicators of sentence boundaries, such as prosody, filled pauses, and discourse markers, can also be affected by cognitive impairments (Emmorey, 1987; Bridges and Van Lancker Sidtis, 2013). For these reasons, it is not clear that

[4] The material presented in this subsection was previously published in: Kathleen C. Fraser, Naama Ben-David, Graeme Hirst, Naida L. Graham, and Elizabeth Rochon (2015). Sentence segmentation of aphasic speech. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT-2015)*, pp. 862–871.

turning to politics for al gore and george w bush another day of rehearsal in just over forty eight hours the two men will face off in their first of three debates for the first time voters will get a live unfiltered view of them together

Turning to politics, for Al Gore and George W Bush another day of rehearsal. In just over forty-eight hours the two men will face off in their first of three debates. For the first time, voters will get a live, unfiltered view of them together.

Figure 4.3: An example of ASR text before and after processing.

existing algorithms can be applied to impaired speech with any degree of success.

Here we explore whether we can apply standard approaches to sentence segmentation to impaired speech, and compare our results to the segmentation of broadcast news. We then extract syntactic complexity features from the automatically segmented text, and compare the feature values with measurements taken on manually segmented text. We assess which features are most robust to the noisy segmentation, and thus could be appropriate features for future work on automatic diagnostic interfaces.

4.2.1 Background: Automatic sentence segmentation

Many approaches to the problem of segmenting recognized speech have been proposed. One popular way of framing the problem is to treat it as a sequence tagging problem, where each interword boundary must be labelled as either a sentence boundary (B) or not (NB) (Liu and Shriberg, 2007).

Liu et al. (2005) showed that using a conditional random field (CRF) classifier for this problem resulted in a lower error rate than using a hidden Markov model or maximum entropy classifier. They stated that the CRF approach combined the benefits of these two other popular approaches, since it is discriminative, can handle correlated features, and uses a globally optimal sequence decoding.

The features used to train such classifiers fall broadly into two categories: word features and prosodic features. Word features can include word or part-of-speech *n*-grams, keyword identification, and filled pauses (Stevenson and Gaizauskas, 2000; Stolcke and Shriberg, 1996; Gavalda et al., 1997). Prosodic features include measures of pitch, energy, and duration of phonemes around the boundary, as well as the length of the silent pause between words (Shriberg et al., 2000; Wang et al., 2003).

The features which are most discriminative to the segmentation task can change depending on the nature of the speech. One important factor can be whether the speech is prepared or spontaneous. Cuendet et al. (2007) explored three different genres of speech: broadcast news, broadcast conversations, and meetings. They analyzed the effectiveness of different feature sets on each type of data. They found that pause features were the most discriminative across all groups, although the best results were achieved using a combination of lexical and prosodic features. Kolár et al. (2009) also looked at genre effects on segmentation, and found that prosodic features were more useful for segmenting broadcast news than broadcast conversations.

Sentence segmentation of written text has also been studied extensively. Recent approaches have reported F-measures of up to 99.8% (Read et al., 2012) and even 100% (Evang et al., 2013). Punctuation provides much of the information in the written case; Read et al. (2012) report that 91.9% of the sentences in their data set end with either a period, question mark, or exclamation mark.

4.2.2 Data

PPA data

For the PPA case, we include 28 patients with PPA (11 with sv-PPA and 17 with nfv-PPA), and 23 age- and education-matched healthy controls.[5] As before, the *Cinderella* narratives were transcribed by trained research assistants, and the transcriptions include filled pauses, repetitions, and false starts. Sentence boundaries were marked by a single annotator according

[5] This data set includes all the narrative samples from Chapter 3, as well as some additional samples that became available since that work was completed. Additionally, all the transcripts had been reviewed and in some cases re-segmented. Therefore, syntactic complexity results reported in this section are not directly comparable to those reported earlier.

to semantic, syntactic, and prosodic cues. We remove capitalization and punctuation, keeping track of original sentence boundaries for training and evaluation, to simulate a high-quality ASR transcript. Of course, a real ASR transcript would contain word recognition errors as well, but we first aim to examine sentence segmentation for impaired speech in the absence of such errors.

Broadcast news data

For the broadcast news data, we use a 804,064 word subset of the English section of the TDT4 Multilingual Broadcast News Speech Corpus[6]. Using the annotations in the transcripts, we extracted news stories only (ignoring teasers, miscellaneous text, and under-transcribed segments). The transcriptions were generated by closed captioning services and commercial transcription agencies (Strassel, 2005), and so they are of high but not perfect quality. Again, we remove capitalization and punctuation to simulate the output from an ASR system.

Since the TDT4 corpus is much larger than our PPA data set, we also construct a small news data set by randomly selecting 20 news stories from the TDT4 corpus. This allows us to determine which effects are due to differences in genre and which are due to having a smaller training set. We refer to this smaller news corpus as TDT4-small to distinguish it from the larger TDT4 corpus.

4.2.3 Methods

Lexical and POS features

The lexical features are simply the unlemmatized word tokens. We do not consider word *n*-grams due to the small size of our PPA data set. To extract our part-of-speech (POS) features, we first tag the transcripts using the NLTK POS tagger (Bird et al., 2009). We use the POS of the current word, the next word, and the previous word as features.

[6] `catalog.ldc.upenn.edu/LDC2005S11`

Prosodic features

To calculate the prosodic features, we first perform automatic alignment of the transcripts to the audio files. This provides us with a phone-level transcription, with the start and end of each phone linked to a time in the audio file. Using this information, we are able to calculate the length of the pauses between words, which we bin into three categories based on previous work by Pakhomov et al. (2010a). Each interword boundary either contains no pause, a short pause (<400 ms), or a long pause (>400 ms).

We calculate the pitch (Talkin, 1995; Brookes, 1997), energy, and duration of the last vowel before an interword boundary. For each measurement, we compare the value to the average value for that speaker, as well as to the values for the last vowel before the next and previous interword boundaries.

We perform the forced alignment using the HTK toolkit (Young et al., 1997). Our pronunciation dictionary is based on the CMU dictionary[7], augmented with estimated pronunciations of out-of-vocabulary words using the "g2p" grapheme-to-phoneme toolkit (Bisani and Ney, 2008). We use a generic acoustic model that has been trained on Wall Street Journal text (Vertanen, 2006).

Boundary classification

We use a conditional random field (CRF) to label each interword boundary as either a sentence boundary (B) or not (NB). We use a CRF implementation called CRFsuite (Okazaki, 2007) with the passive-aggressive learning algorithm. To avoid overfitting, we set the minimum feature frequency cut-off to 20.

To evaluate the performance of our system, we compare the hypothesized sentence boundaries with the manually annotated sentence boundaries and report the F score, where F is the harmonic mean of recall and precision. For the PPA data and the TDT4-small data, we assess the system using a leave-one-out cross-validation framework, in which each narrative is

[7] `www.speech.cs.cmu.edu/cgi-bin/cmudict`

Feature set	**TDT4**	**TDT4-small**	**Controls**	**sv-PPA**	**nfv-PPA**
Chance baseline	0.07	0.07	0.05	0.07	0.06
All	**0.61**	0.57	**0.51**	**0.43**	**0.47**
Lexical+prosody	0.57	0.50	0.44	0.30	0.33
Lexical+POS	0.48	0.36	0.36	0.36	0.40
POS+prosody	**0.61**	**0.59**	0.45	0.39	0.45
POS	0.45	0.39	0.28	0.35	0.39
Prosody	0.50	0.48	0.24	0.23	0.25
Lexical	0.26	0.14	0.18	0.17	0.18

Table 4.5: *F* score for the automatic segmentation method on each data set. Boldface indicates best in column.

sequentially held out as test data while the system is trained on the remaining narratives. For the large TDT4 corpus, we randomly hold out 10% of the corpus as test data, and train on the remaining 90%.

Assessment of syntactic complexity

Once we have segmented the transcripts, we want to assess how the (presumably noisy) segmentation affects our measures of syntactic complexity. Here we consider a number of the syntactic complexity metrics that we have used in previous sections. The metrics are defined in the first column of Table 4.7.

4.2.4 Segmentation results

Comparison between data sets

Table 4.5 shows the performance on the different data sets when trained using different combinations of feature types. We also report the chance baseline for comparison.

We first consider the differences in results observed between the two news data sets. The best results are similar in both groups, although, as would be expected, the larger training sample performs better. However, the difference is small, which suggests that the small size of the PPA data set should not greatly hurt the performance. When we compare the performance of these two groups with different sets of training features, we notice that the difference in performance is greatest when training on lexical features. In a small random sample from the TDT4 corpus, it is unlikely that two stories will cover the same topic, and so there will be little overlap in vocabulary. This is reflected in the results showing that lexical features hurt the performance in this small news sample.

Performance on the news corpus is better than on the PPA data (including the control group). Comparing TDT4-small to the PPA controls, we see that this is not simply due to the size of the training set, so we instead attribute the effect to the fact that speech in broadcast news is often prepared, while in the PPA data sets it is spontaneous.

A closer look at the effect of prosodic features in our training data further shows the difference we observe between prepared and spontaneous speech. When trained on the prosodic features alone, the news data set performs relatively well, while performance on the control data is much worse. These results are consistent with the findings of Kolár et al. (2009) regarding the effect of prosodic features in prepared and spontaneous speech.

When comparing the performance on the control group and on the PPA data, we see that generally, the results are better on the controls. This is to be expected, as the speech in the control group has more complete sentences and fewer dysfluencies. However, it is interesting to note that in many cases, the performance on the nfv-PPA and sv-PPA groups is comparable to the performance on controls. All three data sets achieved the best results when trained with all feature types. This suggests that standard methods of sentence segmentation for spontaneous speech can be effective on PPA speech as well.

Looking at the PPA and control groups with other feature sets, we see that POS features are more important in the nfv-PPA and sv-PPA groups than they are for the control data. A closer look at the transcripts shows us that the PPA participants tend to connect independent clauses with a conjunction more frequently than control participants, and independent clauses are often separated in the manual segmentation. This means that many sentence boundaries in

TDT4-small	**Control**	**sv-PPA**	**nfv-PPA**
PRP_next	long pause	CC_next	long pause
DT_next	*go*	NNS	CC_next
RB	*her*	RB	NN
NNS	NNS	NN	RB_next
long pause	CC_next	RB_next	NNS
pitch<ave	RB	PRP_next	RB
NN	RB_next	energy<ave	short pause
CC_next	PRP_next	RB_prev	PRP_next
energy<ave	IN	VB	no pause
IN_prev	short pause	IN_prev	RB_prev

(a) Features associated with a boundary

TDT4-small	**Control**	**sv-PPA**	**nfv-PPA**
VBD_next	TO_next	*the*	TO_next
the	*so*	PRP$_next	*then*
IN	CC	*and*	*the*
MD_next	NNS_next	*then*	*she*
CC	*the*	VBD_next	VBP_next
VBG_next	*she*	VBZ_next	*and*
VBN_next	*and*	TO_next	*uh*
CD_prev	VBD_next	*'s*	VB_next
a	*of*	*I*	VBD_next
to	*uh*	*a*	*a*

(b) Features associated with a non-boundary

Table 4.6: The 10 features with the highest weights in each CRF model, indicating either that the following interword boundary is or is not a sentence boundary.

the PPA data are marked by conjunctions. This is discussed further in the next section.

When considering the prosodic and lexical feature sets individually, we see that performance is similar in all three cases (control, sv-PPA, and nfv-PPA). However, when we combine prosodic and lexical features together, the performance in the control case increases by a much larger margin than in the two aphasic cases. This suggests that control participants combine words and prosody in a manner that is more predictive of sentence boundaries than in the aphasic case.

Important features

In Table 4.6, we report the 10 features in each data set which are most strongly associated with a boundary or a non-boundary. We consider only the reduced news corpus TDT4-small, for a fair comparison with the PPA data.

The POS tags shown are the output of the NLTK part-of-speech tagger, which uses the Penn Treebank Tag Set (Santorini, 1990). We append '_next' and '_prev' to indicate that this is the POS tag of the next and previous word respectively. Italicized words represent lexical items.

We first consider the features that indicate a sentence boundary (see Table 4.6a). In general, we observe that our minimum frequency cut-off removes many of the lexical features from the top 10. (In the absence of such a cut-off, we observed that very low frequency words can be given deceptively high weights.) The exceptions to this are the words *go* and *her* in the control set. When we look at the data, there are indeed many occurrences of *go* and *her* at the end of sentences, for example, *she was not allowed to go* or *she couldn't go*, and *very mean to her* or *so in love with her*. While these lexical items are not specific to the *Cinderella* story, it seems unlikely that these features would generalize to other story-telling tasks (although we note that the Cinderella story is very widely used in the assessment of aphasia and some types of dementia).

The POS of the given word and its neighbours are generally important features. In all four cases, the next word being a coordinating conjunction or a pronoun is indicative of a boundary. In the three PPA cases, but not the news case, the next word being an adverb is also indicative. Looking at the data, we observe that this very often corresponds to the use of words like *so*, *then*, *well*, *anyway*, etc. This would seem to reflect a difference between the frequent use of discourse markers in spontaneous speech and their relative sparsity in prepared speech.

The POS of the current word is also important. In all cases, a boundary is associated with the current word being an adverb or a noun. In the control data only, the tag IN, representing either a preposition or a subordinating clause, is also associated with a boundary. Although this seems counter-intuitive, an examination of the data reveals that in almost every case, this corresponds to the phrase *happily ever after*. The fact that this feature does not occur in the other PPA groups could indicate that the patients are less likely to use this phrase, but could also be due to our relatively high frequency cut-off.

Another anomalous result is that the tag VB (verb, base form) is associated with a sentence boundary in the sv-PPA case only. Again, examples from the data suggest a probable explanation. In many cases, sentences ending with VB are actually statements about the difficulty of the task, rather than narrative content; e.g., *that's all I can say*, *I can't recall*, or *I don't know*. These statements are consistent with the word-finding difficulties that are a hallmark of sv-PPA.

In the prosodic features, we see that long pauses and decreases in pitch and energy are associated with sentence boundaries in the news corpus. However, the results are mixed in the PPA data. This finding is consistent with our results in Section 4.2.4, and supports the conclusion of Cuendet et al. (2007) and Kolár et al. (2009) that prosodic features are more useful in prepared than spontaneous speech.

We now look briefly at the features which are associated with a non-boundary (Table 4.6b). Here we see more lexical features in the top 10, mostly function words and filled pauses. These features reflect the reasonable assumption that most sentences do not end with determiners, conjunctions, or subjective pronouns. One feature which occurs in the news data but not the PPA data is the next word being a modal verb (MD). This seems to be a result of the more frequent use of the future tense in the news stories (e.g., *the senator will serve another term*), in contrast to the Cinderella stories, which are generally told in the present or simple past tense.

4.2.5 Complexity results

We first compare calculating the syntactic complexity metrics on the manually segmented transcripts and the automatically segmented transcripts. The results are given in Table 4.7. Metrics for which there is no significant difference between the manual and automatic segmentation are marked with "NS". Of course, we do not claim that there is actually no difference between the

Metric	Diff?	Controls		sv-PPA		nfv-PPA	
		Manual	Auto	Manual	Auto	Manual	Auto
Max YD maximum Yngve depth		5.10	4.53	4.45	3.87	4.66	3.83
Mean YD mean Yngve depth		2.97	2.72	2.68	2.44	2.77	2.41
Total YD total sum of the Yngve depths		66.92	53.41	42.48	32.95	49.95	32.57
Tree height average parse tree height		12.56	11.30	10.79	9.81	11.25	9.88
S number of sentences		24.35	31.22	27.73	37.36	18.82	25.47
T number of T-units	NS	31.43	35.13	32.55	39.27	23.29	27.41
C number of clauses	NS	61.48	64.48	57.73	62.45	42.94	46.65
DC number of dependent clauses	NS	24.70	27.30	26.27	26.09	16.59	18.88
CN number of complex nominals	NS	41.39	43.52	38.73	39.64	27.12	27.88
VP number of verb phrases	NS	77.00	79.65	72.09	77.00	51.76	55.24
CP number of coordinate phrases		12.39	10.30	11.55	6.91	7.82	4.18
CT number of complex T-units	NS	14.30	13.52	12.00	11.82	9.29	8.71
MLS mean length of sentence		19.79	16.22	14.04	11.25	15.86	11.60
MLT mean length of T-unit		14.92	13.72	12.19	10.46	12.78	10.66
MLC mean length of clause		7.55	7.21	7.13	6.58	6.89	6.39
T/S T-units per sentence		1.34	1.17	1.15	1.06	1.23	1.08
C/S clauses per sentence		2.64	2.25	1.96	1.70	2.28	1.82
DC/T dependent clauses per T-unit	NS	0.80	0.78	0.73	0.63	0.73	0.69
VP/T verb phrases per T-unit		2.47	2.34	2.11	1.92	2.23	1.98
CP/T coordinate phrases per T-unit		0.40	0.33	0.35	0.17	0.35	0.15
CN/T complex nominals per T-unit	NS	1.32	1.26	1.18	1.10	1.17	1.01
C/T clauses per T-unit		1.99	1.91	1.71	1.58	1.86	1.68
CT/T complex T-units per T-unit	NS	0.46	0.40	0.37	0.32	0.39	0.32
DC/C dependent clauses per clause	NS	0.39	0.41	0.42	0.40	0.38	0.41
CP/C coordinate phrases per clause		0.20	0.17	0.20	0.10	0.19	0.09
CN/C complex nominals per clause	NS	0.65	0.65	0.70	0.71	0.63	0.61

Table 4.7: Mean values of syntactic complexity metrics for the different patient groups. Features which show no significant difference between the manual and automatic segmentation on all three clinical groups are marked as "NS" (not significant).

values, as can be seen in the table, but we use this as a threshold to determine which features are less affected by the automatic segmentation.

All the features relating to Yngve depth and height of the parse trees are significantly different in at least one of the three clinical groups. However, of the eight primary syntactic units calculated by Lu's SCA, six show no significant difference when measured on the automatically segmented transcripts. To examine this effect further, we will discuss how each of the eight is affected by the segmentation process.

Although the number of sentences (S) is different, the number of clauses (C) is not significantly affected by the automatic segmentation, which implies that the boundaries are rarely placed within clauses, but rather between clauses. An example of this phenomenon is given in Example 1:

Manual: And then they go off to the ball and then she comes I dunno how she meets up with this um fairy godmother whatever.

Auto: And then they go off to the ball. And then she comes I dunno how she meets up with this um fairy godmother whatever.

Our automatic method inserts a sentence boundary before the second *and*, breaking one sentence into two but not altering the number of clauses. In fact, the proposed boundary seems quite reasonable, although it does not agree with the human annotator. The correlation between the number of clauses counted in the manual and automatic transcripts is 0.99 in all three clinical groups. The counts for dependent clauses (DC) are also relatively unaffected by the automatic segmentation, for similar reasons.

The T-unit count (T) is also not significantly affected by the automatic segmentation. Since a T-unit only contains one independent clause as well as any attached dependent clauses, this suggests that the segmentation generally does not separate dependent clauses from their independent clauses. This also helps explain the lack of difference on complex T-units (CT). Table 4.7 also indicates that the number of verb phrases (VP) and complex nominals (CN) is not significantly different in the automatically segmented transcripts. Since these syntactic units are typically sub-clausal, this is not unexpected given the arguments above.

The remaining primary syntactic unit, the coordinate phrase (CP), *is* different in the automatic transcripts. This represents a weakness of our method; namely, it has a tendency to insert a boundary before all coordinating conjunctions, as in Example 2:

Manual: So she is very upset and she's crying and with her fairy godmother who then uh creates a carriage and horses and horsemen and and driver and beautiful dress and magical shoes.

Auto: So she is very upset. And she's crying and with her. Fairy godmother who then uh creates a carriage. And horses and horsemen and and driver. And beautiful dress. And magical shoes.

In this case, the manual transcript has five coordinate phrases, while the automatic transcript has only two.

The mean lengths of sentence (MLS), clause (MLC), and T-unit (MLT) are all significantly different in the automatically segmented transcripts. The remaining metrics in Table 4.7 are simply combinations of the primary units discussed above.

4.2.6 Classification of participant groups[8]

Our analysis so far suggests that some syntactic units are relatively impervious to the automatic sentence segmentation, while others are more susceptible to error. However, when we examine the mean values given in Table 4.7, we observe that even in cases when the complexity metrics are significantly different in the automatic transcripts, the differences appear to be systematic. For example, we know that our segmentation method tends to produce more sentences than appear in the manual transcripts (i.e., S is always greater in the automatic transcripts). If we look at the differences across clinical groups, the same pattern emerges in both the manual and automatic transcripts: participants with sv-PPA produce the most sentences, followed by controls, followed by participants with nfv-PPA.

In a classification task, what matters most is not the absolute value of the metric, but the relative differences between groups. Thus we now turn to the question of whether syntactic complexity metrics derived from the automatically segmented transcripts are useful in a classification framework, even though their values are different from those derived from manually segmented transcripts.

We perform two classification tasks, PPA versus controls and sv-PPA versus nfv-PPA. We once again use a p-value filter, restricting the size of the feature set to $N = 10$. Features are limited to those given in Table 4.7. We consider three classification algorithms: naïve Bayes

[8] The classification experiment described in this section did not appear in the original paper.

	Auto	Manual
SVM	**.71**	.65
LR	**.75**	.65
NB	**.76**	.69

(a) PPA versus controls

	Auto	Manual
SVM	**.64**	.57
LR	.61	**.68**
NB	**.68**	.50

(b) sv-PPA versus nfv-PPA

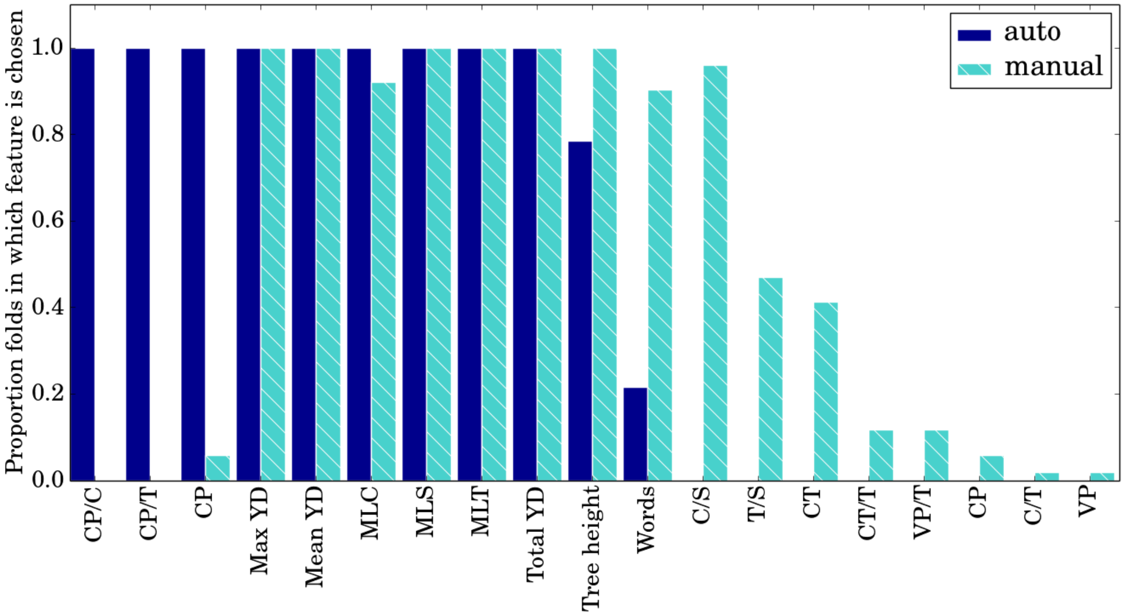
Table 4.8: Classification accuracies for the automatically and manually segmented transcripts. The majority class baseline is .55 for PPA versus controls and .61 for sv-PPA versus nfv-PPA.

(NB), logistic regression (LR), and support vector machines (SVM). The classifiers are trained and tested in a leave-one-out cross-validation procedure, and average accuracy across folds is computed.

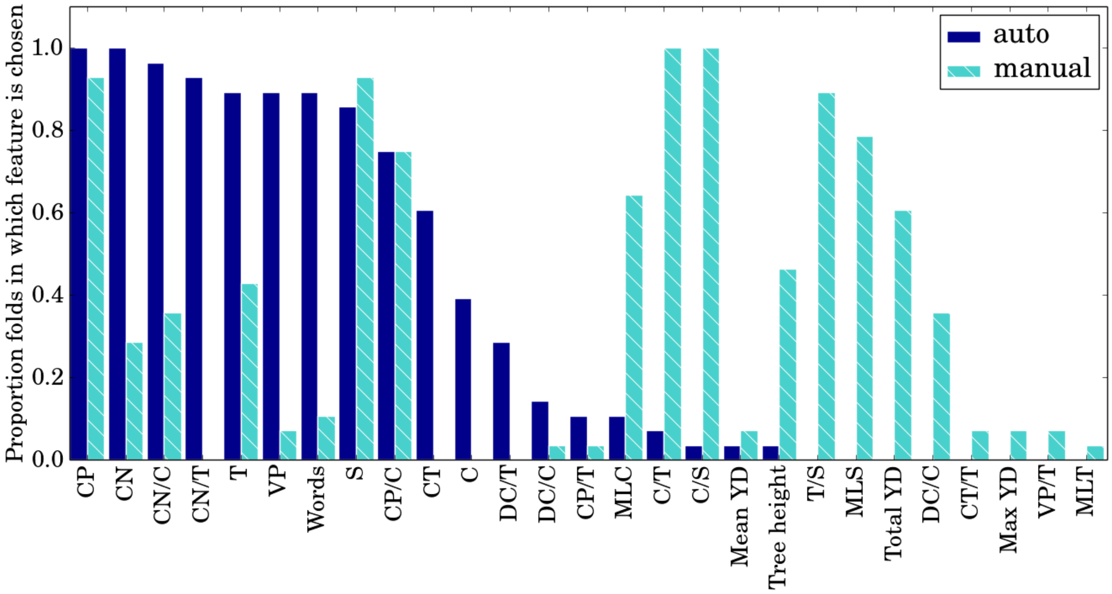
The results of this experiment are shown in Table 4.8. In all cases except one, the best result is actually achieved using the automatically segmented transcripts. To investigate this result further, we examine the features selected by the feature selection step in each case. Since different features may be selected in each fold, for each feature we measure the proportion of folds in which it is selected. This information is shown in Figure 4.4.

In the case of controls versus PPA, some features are selected on both sets of transcripts, but there is also notable divergence. Coordinate phrases (CP) appear to be much more relevant to the distinction between groups in the automatically segmented transcripts. The features CP, CP/C, and CP/T are selected in every fold for the auto transcripts, but very rarely in the manual transcripts. From Table 4.7, we know that the number of coordinate phrases is reduced in the auto transcripts, and the reduction is greater in the patient transcripts (17% reduction in the control group compared to 40% and 47% in the sv-PPA and nfv-PPA groups, respectively). In general, this reduction is most likely due to the segmentation classifier's tendency to split sentences before the word *and*, as discussed above. A possible explanation for why this error affects the patient groups more than the control group is that the controls use more typical prosody, and so the acoustic features predicting "not a boundary" outweigh the lexical features predicting "boundary".

In contrast, features such as clauses per sentence (C/S) are selected highly frequently in



(a) Features selected in the PPA vs. controls task.



(b) Features selected in the sv-PPA vs. nfv-PPA task.

Figure 4.4: Selected features for the two classification experiments, using automatically and manually segmented transcripts.

the manual transcripts and never in the auto transcripts. Upon closer examination of the data, we notice an obvious outlier in the control group on this measure in the manual transcripts. A sample sentence from this transcript reads:

> I believe the um the fairy godmother told her she had to be home by twelve so she starts to see that it's getting nearer and she gets worried and she starts running down the stairs and she loses one of her slippers uh she had a glass slipper on and anyway she goes home and then uh the prince is wondering what happened to that girl because I guess he fell for her and um only in the fairy tales so anyway then they send someone from the palace with this glass slipper to s find the girl who fits in the glass slipper and of course the the wicked stepsisters it doesn't fit them but then it fits Cinderella and she gets her prince and they live happily ever after.

The same passage was segmented by the automated method into the following sentences:

> I believe the um the fairy godmother told her she had to be home by twelve.
>
> So she starts to see that it's getting nearer and she gets worried and she starts running down the stairs and she loses one of her slippers uh she had a glass slipper on and anyway she goes home and then uh the prince is wondering what happened to that girl because I guess he fell for her.
>
> And um.
>
> Only in the fairy tales so anyway then.
>
> They send someone from the palace with this glass slipper to s find the girl who fits in the glass slipper and of course the the wicked stepsisters it doesn't fit.
>
> Them but then it fits Cinderella and she gets her prince and they live happily ever after.

Clearly, the automated method is not perfect (e.g., the first word of the last sentence belongs with the previous sentence), but it seems fair to say that the human segmentation is not perfect, either. Importantly, the manual segmentation of this particular narrative is not consistent with how the other narratives were segmented, leading to anomalous results on the syntactic complexity measures that do not necessarily reflect real differences between the groups.

One positive characteristic of the features chosen on the auto transcripts is that there is very little variation between folds. Nine of the 11 features are selected in 100% of the folds. This indicates that the features measured on the training set are likely to generalize well to the test set (as reflected in the higher classification accuracies). Furthermore, although the values of the features differ from those measured on the manual transcripts, the direction of the trend is the same as in the manual data set on all 11 selected features.

In the case of sv-PPA vs. nfv-PPA, there is more variation in the selected features for both the auto and manual transcripts. The classification results are also lower. In this case, complex nominals (CN; also CN/C and CN/T) are chosen frequently in the auto transcripts and rarely in the manual transcripts. Interestingly, complex nominals were one feature that was found to be robust to the noisy segmentation. However, if we rank the features by p-value in both the auto and manual data sets, we find that CN is ranked first in the auto case ($p = .13$) and tenth in the manual case ($p = .19$). That is, the feature does not significantly differ between sv-PPA or nfv-PPA in *either* group, it is simply ranked higher in the auto group. Given that there are in fact no significant features in the auto transcripts and only one (C/S) in the manual transcripts, we will not attempt to provide further interpretation for these features.

4.2.7 Discussion

We have examined the issue of automated sentence segmentation of impaired speech, and tested the effectiveness of standard segmentation methods on PPA speech samples. We found that, as expected, performance was best on prepared speech from broadcast news, then on healthy controls, and worst on speech samples from PPA patients. However, the results on the PPA data are promising, and suggest that similar methods could be effective for aphasic speech. Future work will look at adapting the standard algorithms to improve performance in the case of impaired speech. This would include an evaluation of the forced alignment on impaired speech data (e.g., by comparing the labelled phone start and end times with manually annotated transcripts), as well as the exploration of new features for the boundary classification.

One limitation of this study is the use of manually transcribed data with capitalization and punctuation removed to simulate perfect ASR data. We expect that real ASR data will contain recognition errors, and it is not clear how these errors will affect the segmentation process. Unfortunately, given the extremely high WERs reported in the previous section, it is not possible to determine the location of the gold standard sentence boundaries in those ASR transcripts. Analysis of boundary segmentation in real ASR data will have to wait until we have more accurate ASR for this population.

We analyzed our results to see how the noise introduced by our segmentation affects various syntactic complexity measures, by comparing the values of these measures computed on the automatically and manually segmented transcripts. Some measures (e.g., T-units) were robust to the noise, while others (e.g., Yngve depth) were not. When using such automatic methods for the analysis of speech data, researchers should be aware of the unequal effects on different complexity metrics.

We then tested the effectiveness of the syntactic measures on two classification tasks, and found that in most cases the accuracy was actually higher using the automatically segmented transcripts than the manually segmented transcripts. While this result was surprising, we identified two possible factors that may have contributed to this effect: segmentation errors that preferentially affected the PPA group, and more consistent segmentation within groups in the automated case.

Although we evaluated our methods against human-annotated data, there is some uncertainty about whether a single gold standard for the sentence segmentation of speech truly exists. Miller and Weinert (1998), among others, argue that the concept of a sentence as defined in written language does not necessarily exist in spoken language. In future work, it would useful to compare the inter-annotator agreement between trained human annotators to determine an upper bound for the accuracy of an automatic system.

4.3 Summary

In this chapter, we explored the question of whether we can apply ASR techniques to fully automate our processing pipeline. Despite the growing popularity of speech-based interfaces, there is still much work to be done before wide-vocabulary, speaker-independent speech recognition will reach an acceptable level of accuracy. In the meantime, however, we did achieve some promising results. In Section 4.1, we showed that even when the WER is very high, psycholinguistic features like frequency, familiarity, and imageability can still help to distinguish between PPA patients and controls. In ongoing work, we are exploring the effect of adapting the acoustic and language models to the specific parameters of our data set, using open-source ASR toolkits such as HTK and Kaldi.

A fully automated system will also require automatic utterance segmentation. In Section 4.2, we applied standard sentence segmentation algorithms to PPA speech. While the results were not as good as the results achieved on broadcast news data, this was to be expected due to the nature of spontaneous speech. Segmentation of PPA speech was also less successful than segmentation of control speech, and we hypothesize that this is at least partly due to the increased incidence of pausing within sentences. However, we found that several of the syntactic features were robust to the noisy segmentation, and in fact better classification accuracies were achieved using the automatically segmented transcripts.

Somewhat separate from the specific issues of ASR and segmentation, the results in this chapter raise important questions about the goal of feature selection and machine learning in the context of medical diagnostics and monitoring. Optimizing the accuracy of the prediction is the primary (and in some cases only) goal in many machine learning applications. However, as Guyon and Elisseeff (2003) point out, this is not the only role of feature selection in bioinformatics problems. Ideally, we want to look to the set of selected features to inform our knowledge of the health condition at hand. Certainly an application that outputs a probability of diagnosis but no report of the underlying factors which led to the prediction is unlikely to be useful in a clinical context. Thus, while the classification results presented in this chapter are promising, further work is required to refine our understanding of these “noisy” variables and their relationship to ground truth measurements.

Chapter 5

Agrammatism in progressive and post-stroke aphasia

It is an open question whether the grammatical impairments seen in primary progressive aphasia (PPA) are similar in nature to those seen in post-stroke agrammatic aphasia, or if instead there are meaningful differences between the two conditions. From a purely anatomical perspective, there is no reason why the two conditions must have the same linguistic presentations: damage to the brain after a stroke tends to follow a distinct pattern, defined by the structure of the cerebral vascular system, while damage due to neurodegenerative disease does not (Thompson et al., 2013). Yet, in many cases, the deficits seen in nonfluent PPA are similar to those reported in post-stroke Broca's aphasia (Rohrer et al., 2008; Harciarek and Kertesz, 2011).

In this preliminary work, we show that computational analysis may provide insight on this debate. By extracting linguistic features from the narrative speech of agrammatic participants with PPA and post-stroke aphasia, we can train a classifier to distinguish between the two groups with 76% accuracy, suggesting that there are some quantifiable differences between the groups. The highly-ranked features are compared to manually extracted measures and discussed in relation to previous work in the field.

5.1 Background

Agrammatism is a common symptom in Broca's aphasia (Goodglass, 1993), and can manifest in a variety of different ways. Wilson et al. (2012) define grammatical processing as "the ability to implicitly generate hierarchically structured representations of sentences, and to use function words and inflectional morphology to express grammatical categories such as number, definiteness, tense and aspect." There are a number of different linguistic concepts wrapped up in this definition, and studies have shown dissociations between the various components (Saffran et al., 1989; Webster et al., 2007; Nadeau, 2012; Webster and Howard, 2012). Thompson and Mack (2014) broadly break down this idea of grammar into the following elements: grammatical morphology, functional categories (i.e., parts-of-speech or POS), verb-argument structure, and syntactic complexity. We will briefly summarize the findings in these four categories with respect to speech production in agrammatic aphasia.

Grammatical morphology refers to the modifications made to words to express their relationship to other words, such as agreement in tense, number, aspect, and gender. In agrammatic aphasia, these grammatical morphemes tend to be either omitted or mis-used (Thompson and Bastiaanse, 2012; Webster et al., 2007). With respect to the production of different POS, people with agrammatic aphasia show a marked reduction in the production of verbs, relative to nouns (Saffran et al., 1989; Kim and Thompson, 2000; Goodglass et al., 1994). They also have a tendency to omit function words (or closed-class words) relative to content words (or open-class words) (Saffran et al., 1989; Thompson et al., 1995). The verb-argument structure of a verb defines the constraints placed on a verb's arguments, including subcategorization (e.g., an intransitive verb can have only a subject, whereas a ditransitive verb requires both a subject and two objects) and selectional constraints (e.g., the verb *eats* requires an animate subject). People with agrammatic aphasia tend to produce verbs with fewer arguments, and may make errors in the argument structure (Thompson et al., 1995; Kim and Thompson, 2000). Finally, the speech output of agrammatic patients tends to show a reduction in syntactic complexity (Saffran et al., 1989; Goodglass et al., 1994).

Agrammatism in language production is also one of the two core criteria for nonfluent/agrammatic PPA (nfv-PPA) (Gorno-Tempini et al., 2011). However, since only one of the two core criteria need be satisfied, it is possible for patients to be diagnosed with nfv-PPA without showing signs of agrammatism. This has led to some disagreement in the literature regarding the nature and extent of agrammatism in nfv-PPA (Thompson et al., 1997a; Graham et al., 2004; Knibb et al., 2009; Wilson et al., 2010; Thompson et al., 2012, 2013; Thompson and Mack, 2014).

Very few studies have directly compared the speech of people with agrammatism due to PPA and due to stroke. However, two previous studies are particularly relevant to this work. Patterson et al. (2006) compared the performance of 10 participants with nfv-PPA and 10 participants with post-stroke nonfluent aphasia (NFA) on a number of language tasks, including picture description, reading passages and lists, and tests of phonological judgement and manipulation. Their hypothesis was that the deficits seen in nfv-PPA would be more sensitive to the specific language task, and more specifically, that these deficits would primarily appear in tasks requiring "self-generated" connected language. Indeed, they found that while there was no difference between the groups on speech rate for the picture description task, the average nfv-PPA speech rate nearly doubled on the reading task, while the average NFA speech rate stayed the same. They also found that nfv-PPA patients made roughly half the amount of errors reading function words in context that NFA patients did (errors on function words in spontaneous speech were not reported).

Patterson et al. (2006) also report a pattern, confirmed by previous literature, that NFA patients experience the most difficulty reading non-words, then function words, and then low-imageability words, with high-imageability words presenting the least difficulty. The nfv-PPA patients (a) performed better than NFA patients on reading each of these categories of words, and (b) showed minimal differences across each of the categories. The authors conclude that spontaneous connected speech is specifically impaired in nfv-PPA (in contrast to NFA, in which participants showed equivalent deficits in spontaneous and read speech), and that phonological processing capabilities are better in nfv-PPA than NFA.

Thompson et al. (2013) presented results for a number of different experiments comparing syntactic processing in two subtypes of PPA and in post-stroke agrammatic and anomic aphasia. Specifically, the experiment examining narrative discourse (via a story-telling task) included 9 participants with agrammatic PPA (PPA-G) and 8 participants with post-stroke agrammatic aphasia (StrAg). There was no significant difference between the two groups on speech rate, mean length of utterance, proportion of grammatical sentences, proportion of correctly inflected verbs, or noun-verb ratio. However, there was a significant difference on the ratio of open-class words to closed-class words, with the StrAg group having a higher value. When compared to healthy controls, the StrAg open-closed class ratio was significantly higher, while there was no difference between PPA-G and controls. This suggests that PPA-G patients do not show a deficit in producing closed-class (function) words. Thompson et al. (2013) note that this result is consistent with previous work by Graham et al. (2004), but disagree with the interpretation given by Graham et al. that nonfluent PPA patients do not show true agrammatism. In fact, the authors conclude that despite the different disease processes underlying PPA-G and StrAg, the resulting language deficits are very similar.

There are a few possible explanations for the apparent disagreement between the results reported by Patterson et al. (2006) and Thompson et al. (2013). One factor is that the two studies measure disjoint sets of variables, apart from speech rate (in which case, both studies report no significant difference between the groups for spontaneous, connected speech). Another factor is the relatively small sample sizes included in each study. However, possibly the most salient factor is related to the classification of PPA subtypes. In Patterson et al. (2006), the PPA participants are described as having *progressive nonfluent aphasia* and exhibiting “slowed, effortful output with phonological errors in spontaneous speech or on formal testing.” However, the participants in Thompson et al. (2013) were subtyped according to a different set of criteria, which allows only those participants with frank grammatical impairments into the PPA-G class. It is therefore possible that Thompson et al. (2013) observe grammatical deficits in their PPA group specifically because agrammatism was a diagnostic criterion for inclusion in the group, while Patterson et al. (2006) do not observe as much frank agrammatism (and nor do Graham et al. (2004) in a companion study using the same participants) because their PPA patients were selected on the basis of "nonfluency" rather than agrammatism. Furthermore, since the Patterson et al. (2006) paper was published before the consensus criteria (Gorno-Tempini et al., 2011), it is also possible that their "nonfluent" group includes participants who today would be classified as lv-PPA.

In the work which follows, patient diagnoses were made according to the same criteria as in the work of Thompson et al. (2013). Thus we know that both participant groups exhibit agrammatism, and the question becomes: is there any difference in the nature of the agrammatism, or in other linguistic areas (e.g., semantic processing), between the progressive and acute conditions?

5.2 Data

This study is a retrospective analysis of data that was collected by Cynthia Thompson and colleagues at Northwestern University. Participants were recruited from subject pools at the Northwestern University Aphasia and Neurolinguistics Research Laboratory. All participants were native speakers of English. Demographic information is given in Table 5.1. A diagnosis of PPA was made by an experienced neurologist on the basis of a neurological examination, neuropsychological testing, and clinical presentation. PPA subtyping was based on single word comprehension and sentence generation tasks. Specifically, a diagnosis of PPA-G was given when single word comprehension was spared, but sentence generation was impaired. Only those participants with a primarily grammatical impairment (as opposed to a motor speech impairment) were included in this study. The StrAg participants had developed aphasia as the result of a left-hemisphere stroke, and were classified as agrammatic on the basis of formal language testing and clinical impression.

	PPA-G ($n = 35$)	StrAg ($n = 27$)	sig?
Age (years)	64.1 (6.9)	55.4 (11.8)	$p < 0.01$
Education (years)	16.1 (2.6)	16.3 (2.4)	n.s.
Sex (M/F)	20/15	17/10	n.s.
WAB-AQ	82.5 (6.9)	78.6 (8.9)	n.s.

Table 5.1: Demographic information for PPA-G and StrAg participants.

All participants completed the Western Aphasia Battery (WAB) (Kertesz, 1982). The WAB Aphasia Quotient (WAB-AQ) provides an assessment of overall aphasia severity. Although the patient groups originally included 39 participants with PPA-G and 28 participants with StrAg, to match the groups for aphasia severity, it was necessary to remove the 4 PPA-G participants with the highest WAB-AQ and the StrAg participant with the lowest WAB-AQ. There is no significant difference between the resultant groups on aphasia severity. There is also no significant difference on education or sex. There is a significant difference between the ages, with the PPA-G group being older on average. This is due to the fact that people with neurodegenerative diseases tend to be older, in general, and a significant age difference between groups was also reported in Patterson et al. (2006) and Thompson et al. (2013).

Narrative speech data was elicited through a Cinderella story-telling task. Participants were given a wordless picture book to remind them of the story, then the book was taken away and they were asked to tell the story in their own words. The narratives were recorded and then manually transcribed. Dysfluencies such as repetitions, filled pauses, and comments on the task were manually annotated and removed.

5.3 Methods

Measures of linguistic performance were both manually and automatically extracted from the Cinderella transcripts. These features were used to train a machine learning classifier to distinguish between the two patient groups. Details of the features extraction and classification procedure are given in the following sections.

5.3.1 Manually calculated features

As part of previous and ongoing work by the Northwestern University Aphasia and Neurolinguistics Research Lab, each participant narrative has been manually annotated for syntax and morphology by trained annotators, following the NNLA procedure. A number of different measures were calculated from these annotations by researchers at Northwestern. These manually computed features are given in Table 5.2.

Some of the features in Table 5.2 involve a Verb Morphology Index (VMI). This is a measure of the morphological complexity of a verb that was developed by Thompson et al. (1995) and adapted from a scoring system originally proposed by Chomsky (1957). The VMI increases in the presence of tense markings, auxiliary verbs, modal verbs, etc.

5.3.2 Automatically extracted features

As discussed in Section 3.4, we had previously examined the usefulness of the CFG features for distinguishing between speakers with agrammatic aphasia and healthy controls (Fraser et al., 2014b). In the aphasic case, we found evidence for fragmented speech with a higher incidence of solitary noun phrases, difficulty with determiners and possessives, and a reduced number of prepositional phrases and embedded clauses. Furthermore, using these features in a classification task led to better results than using traditional measures of syntactic complexity. Given that we have shown these features to be sensitive to agrammatism, we now examine whether they, in combination with the other text features, can actually distinguish between the two different agrammatic populations.

Text-based features were extracted from the transcripts following the procedures outlined in Chapter 3. In addition, we introduce a new set of features derived from the dependency structure of each narrative utterance, described below. Acoustic features were not included.

Feature name and description
MLU-word Mean length of utterance in words.
MLU-morpheme Mean length of utterance in morphemes.
WPM Number of words per minute (excluding dysfluencies).
TTR Type-token ratio.
Gram-sentences Percentage of sentences which are grammatically correct.
Syn-sentences Percentage of sentences which are syntactically correct.
Complexity-ratio Ratio of complex sentences to simple sentences (where a simple sentence is produced in canonical form with no embedded clauses).
Open-closed-ratio Ratio of number of open-class words to number of closed-class words.
Noun-verb-ratio Ratio of number of nouns to number of verbs.
Correct-reg-infl Percentage of regular verbs which are correctly inflected.
Correct-irreg-infl Percentage of irregular verbs which are correctly inflected.
Correct-3s Percentage of third-person singular verbs which are correctly inflected.
Correct-ed Percentage of past-tense verbs which are correctly inflected.
Correct-comp Percentage of complementizers which are used correctly.
Percent-3s Percentage of third-person singular markers that are produced, out of the total number of regular morphological markers.
Percent-ed Percentage of regular past-tense markers that are produced, out of the total number of regular morphological markers.
Percent-ed+ Percentage of regular past tense and past participle markers that are produced, out of the total number of regular morphological markers.
Percent-ing Percentage of present participle markers that are produced, out of the total number of regular morphological markers.
Correct-1-place Percentage of 1-place verbs with the correct argument structure.
Correct-2-place Percentage of 2-place verbs with the correct argument structure.
Correct-3-place Percentage of 3-place verbs with the correct argument structure.
Correct-arg-struct Percentage of verbs with the correct argument structure.
Correct-VMI Percentage of verbs with correct Verb Morphology Index (VMI).
Mean-VMI Mean VMI (calculated over all verbs).
Correct-mean-VMI Mean VMI (calculated over correct verbs).
Correct-VI Percentage of correctly inflected verbs.
Correct-NI Percentage of correctly inflected nouns.

Table 5.2: Features which were manually extracted from the narrative samples in previous work at Northwestern University.

Dependency parse features

The syntactic features used throughout this thesis have been extracted from context-free grammar constituent parses of sentences. A different grammar formalism, called a dependency grammar, describes a sentence in terms of the relationships between words, rather than in terms of the syntactic constituents (Jurafsky and Martin, 2000). In dependency parsing, the nodes are words and the edges are the grammatical relationships between words (e.g., subject, object, modifier).

One reason for limiting our previous analysis to constituent parse features (and in that case, to ignore the lexical productions) was to mitigate issues of data sparsity when calculating the features. For example, in the sentence *Cinderella washed the laundry and swept the floor*, we can count two instances of the rule VP → VBD NP and two instances of the rule NP → DT NN, but it is not obvious how (or whether) to conflate the associated dependency relations, when the lexical content and sentence positions are different:

```
nsubj(washed-2, Cinderella-1)
root(ROOT-0, washed-2)
det(laundry-4, the-3)
dobj(washed-2, laundry-4)
cc(washed-2, and-5)
conj(washed-2, swept-6)
det(floor-8, the-7)
dobj(swept-6, floor-8)
```

Instead, as a first step towards including some of this relational information, we simply count the frequency of each relation type, without considering the arguments. In the example above, the counts would be *nsubj*: 1, *root*: 1, *det*: 2, *dobj*: 2, *cc*: 1, *conj*: 1. This allows us to capture rough estimates for some of the quantities that had been calculated by hand; for example, the number of direct objects provides an estimate for the number of verbs with greater than one argument.

Dependency parses are obtained from the Stanford dependency parser[1] (de Marneffe et al., 2006). In related work, da Cunha et al. (2015) used dependency parsing to automatically identify propositions in order to calculate the *idea density* of speech from participants with dementia. Roark et al. (2011) used mean dependency distance as a feature when classifying MCI narratives from controls. Orimaye et al. (2014) computed the number of unique dependencies and the average number of unique dependencies per sentence in their paper on language in Alzheimer's disease, although they found no difference on those measures relative to healthy controls. More broadly related work has considered using dependency parse features to track child language development (Lubetich and Sagae, 2014) and to detect preposition errors in second-language writing (Tetreault et al., 2010).

5.3.3 Classification

We used the extracted features to train a logistic regression classifier and a support vector machine classifier (Hall et al., 2009) for varying feature set sizes. A correlation-based feature selection method was used, in which features were ranked according to their correlation with diagnosis (Guyon and Elisseeff, 2003). Performance was measured using leave-one-out cross-validation and calculating the average accuracy across the entire data set. Error bars were estimated based on a 95% confidence interval using the normal approximation method. A majority-class classifier would achieve an accuracy of 0.56.

[1] Here we also go back to using the Stanford constituency parser, for consistency. While we did not compare the results using the Charniak and Stanford parsers, it would be an interesting avenue for future research.

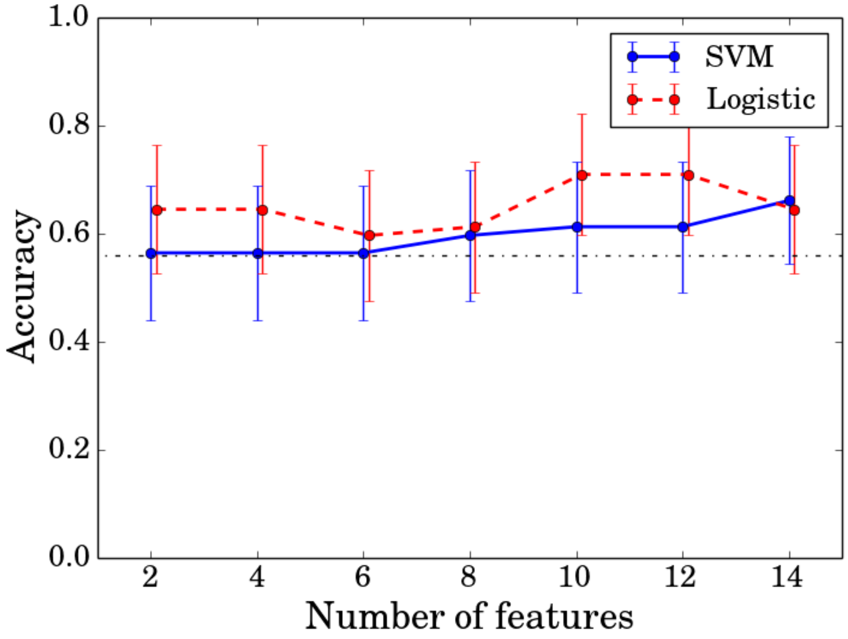


Figure 5.1: Classification accuracy for various feature set sizes using the manually extracted features. The majority-class baseline is shown with a black dotted line.

5.4 Results

5.4.1 Classification results

The classification accuracies achieved using the manual features are given in Figure 5.1. Performing the classification revealed one potential drawback to using these features, which is that the feature values are frequently undefined (for example, the percentage of 3-place verbs with correct argument structure when no 3-place verbs are used). Features with unknown values are not selected in the feature selection step.[2] Thus, the maximal feature set size for the manual features on this data set is 14. The maximum classification accuracy is 0.71 (logistic regression, $N = 10, 12$). This is significantly greater than the baseline (paired t-test, $p = 0.04$).

The accuracies achieved using the automatically extracted features are given in Figure 5.2. Here, the best accuracy is 0.76 (logistic regression, $N = 70$). This accuracy is higher than

[2] When the feature selection was modified to allow for this, a drop in performance was observed.

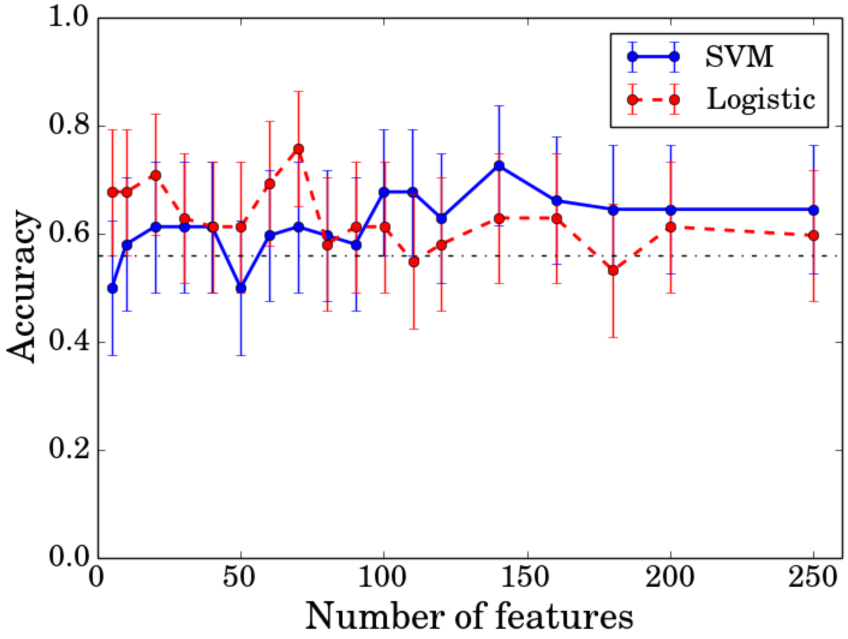


Figure 5.2: Classification accuracy for various feature set sizes using the automatically extracted features. The majority-class baseline is shown with a black dotted line.

using the manual features, but the difference is not significant. It is significantly higher than the majority-class baseline (paired t-test, $p = 0.01$).

5.4.2 Analysis of important features

We now turn to the question of what information the classifiers are using to distinguish between the groups in the automated analysis. In the feature selection step, those features with the highest correlation with diagnosis are selected. These features can be slightly different from fold to fold, as the training sets are different from fold to fold. Figure 5.3 shows which features are selected most frequently across folds (limited to the $N = 10$ case for the sake of visualization and brevity). There are 5 features which are selected in every fold: words per minute (WPM), VP proportion (the length of each verb phrase in words, divided by total narrative length in words), the number of *aux* (auxiliary) dependency relations, the frequency of occurrence of the grammatical structure VP → VBG S, and the proportion of light verbs. A few other features are selected more than 50% of the time: the frequency of NP → NN NN, the frequency of the relation *prep on* (prepositional relation using *on*), PP rate (the number of prepositional phrases, divided by the number of words in the narrative), and the frequency of NP → NNP.

To explore these highly-ranked features in more depth, the means of the 9 features mentioned above in both the PPA-G group and the StrAg group are reported in Table 5.3. WPM is significantly higher in the PPA-G group ($p = 0.001$), in contrast to previous studies that reported no difference between the groups on speech rate (Patterson et al., 2006; Thompson et al., 2013). The PPA-G group also has a higher proportion of words which are part of a verb phrase (note that this proportion can be greater than one because each word can be part of multiple, nested verb phrases). Although not selected in the top features, average VP length is also higher in the PPA-G group (5.0 versus 4.3, $p = 0.03$). People with agrammatic aphasia may omit words like auxiliaries from verb phrases (Saffran et al., 1989), produce fewer verbs with complex argument structures (Thompson et al., 1997a), and use fewer adjectives (Nadeau, 2012), which could all contribute to shorter VPs. Thus it would appear that the PPA-G group is less impaired than the StrAg group on this measure.

The *aux* feature is higher in the PPA-G group than the StrAg group, and this may be related to the VP proportion feature. Consider the following examples:

```
Cinderella cleans the kitchen.
(ROOT
  (S
    (NP (NNP Cinderella))
    (VP (VBZ cleans)
      (NP (DT the) (NN kitchen)))
    (. .)))
```

This sentence has one verb phrase containing 3 words, and a total length of 4 words, so the VP proportion is 3/4 = 0.75.

```
Cinderella is cleaning the kitchen.
(ROOT
  (S
    (NP (NNP Cinderella))
    (VP (VBZ is)
      (VP (VBG cleaning)
        (NP (DT the) (NN kitchen))))
    (. .)))
```

This sentence has one VP containing 3 words, and another VP containing 4 words, so the total VP proportion is (3+4)/5 = 1.4. It also contains the auxiliary verb relation *aux(cleaning-3, is-2)*. So it is clear that sentences with auxiliary verbs will also tend to have higher VP proportions. Again, because use of auxiliary verbs is impaired in agrammatic aphasia, this suggests a lower degree of impairment in PPA-G.

The feature VP $\rightarrow$ VBG S is also higher in the PPA-G group. The means in Table 5.3 are normalized by the total number of CFG rules produced, but on average PPA-G participants use this construction 1.1 times per narrative, while StrAg participants use this construction only 0.44 times per narrative. When we examine the PPA-G data, we find that this construction corresponds to two main usages: (1) Past progressive with infinitive complement. Examples: *she was trying to dance*, *he was trying to find the owner of the shoe*. (2) Future progressive using *going to*. Examples: *the prince is going to find the lady*, *she is going to go to the party*. In each of these cases, the auxiliary verb *be* is used (*was* trying, *is* going to), as well as the progressive (*-ing*) verb form. In general, progressive verb forms are over-used in agrammatic aphasia (Druks and Carroll, 2005; Faroqi-Shah and Thompson, 2007), but auxiliaries tend to be omitted (Thompson and Bastiaanse, 2012).

Table 5.3 shows a higher proportion of light verbs in the PPA-G group. Given the evidence so far for PPA-G participants using more auxiliary verbs, one explanation for the increased number of light verbs in the PPA-G narratives could be that the auxiliary verbs are being labelled as light verbs (since the tagger does not distinguish between auxiliary verbs and regular verbs). If the effect is real, then the interpretation is still unclear. Previous studies do not agree on whether agrammatic patients tend produce more light verbs than controls (as reported by Berndt et al. (1997)) or fewer (as reported by Breedin et al. (1998)).

The feature NP $\rightarrow$ NN NN occurs more frequently in the StrAg group (on average 0.48 times per narrative in the StrAg group versus only 0.03 in the PPA-G group). While there are some grammatical uses of this construction, it often corresponds to an ungrammatical production, e.g., *man coach riding* and *pearl earring were broken*.

The PP rate is higher and prepositional relations with the preposition *on* are more common in the PPA-G group. In general, people with agrammatism tend to produce fewer prepositional phrases (Goodglass et al., 1994). In particular, processing of locative prepositions (such as *in* or *on*) can be impaired in agrammatic aphasia (Nadeau, 2012). However, there is some question of whether the effect here can be interpreted as a true difference in the production of prepositional phrases. Examining the PPA-G data, we find that while many of the examples of *prep on* do correspond to the preposition *on* (e.g., *the slipper she left on the stair*, *it fit on her foot*), there are also many examples which may be more appropriately marked as phrasal verbs (e.g., *she put on the shoe*, *having the people try on the slipper*). Thompson et al. (1995) found that people with nonfluent aphasia tended to produce fewer phrasal verbs than healthy controls in conversation. However, it will be necessary to develop new features to more accurately examine the use of *on* as both a preposition and a particle in a phrasal verb separately before we can draw a conclusion here.

The last feature in Table 5.3 relates to the construction NP $\rightarrow$ NNP, which unsurprisingly corresponds to the word *Cinderella* in almost all cases. On average, the StrAg group uses this construction 4.9 times per narrative, while the PPA-G groups uses it 3.0 times per narrative. Since there is no difference on the total number of words produced, one potential explanation could be that people with PPA-G are referring to Cinderella with a pronoun more frequently. Since *she* could refer to Cinderella, the fairly godmother, the step-mother, or one of the step-

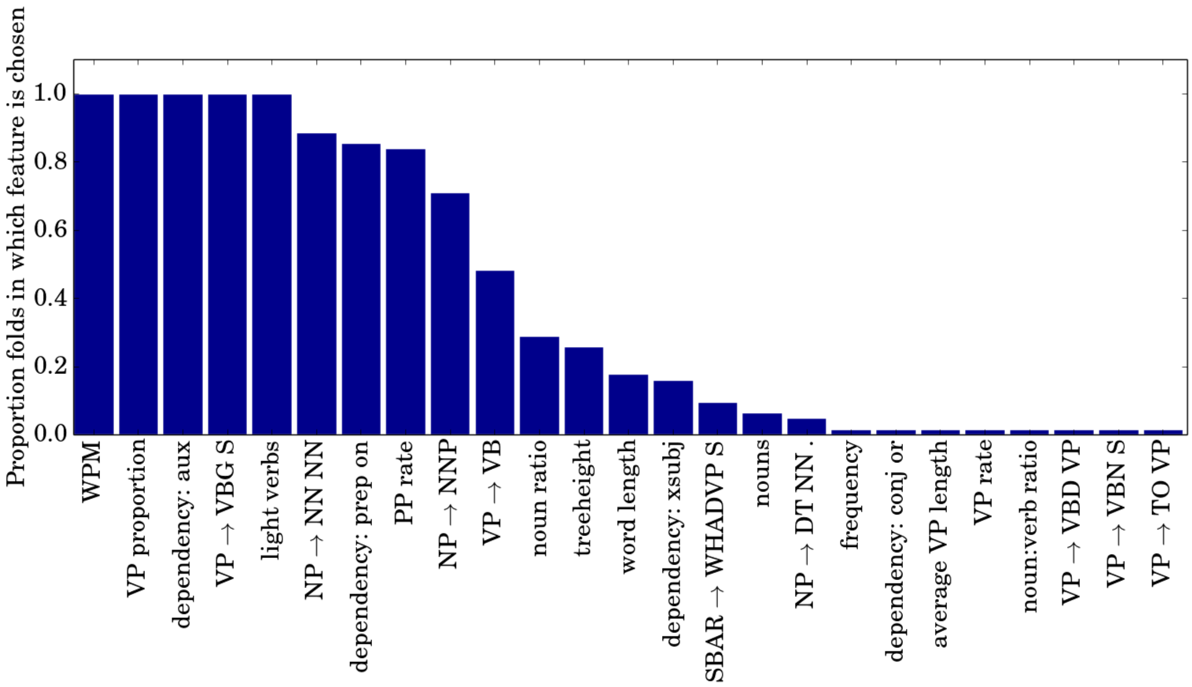


Figure 5.3: The proportion of folds in which each feature was selected, for the case of $N = 10$.

sisters, it is difficult to test this hypothesis directly without performing co-reference resolution, but as an indirect measure we can calculate the ratio of proper nouns to pronouns in each narrative. Indeed, we find that on average the PPA-G group ratio is lower (0.22 versus 0.62 for the StrAg group, $p = 0.04$). The "pronoun ratio" feature, which measures the ratio of pronouns to nouns+pronouns, is not significantly different between the groups (although slightly higher in the PPA-G group). If the PPA-G group is using more pronouns, one explanation could be a subtle word-finding difficulty. Alternatively, it may reflect a difficulty with pronouns in the StrAg group. In previous work, participants with nonfluent PPA showed spared production of pronouns relative to healthy controls (Wilson et al., 2010), while participants with StrAg showed a reduction in pronoun production (Saffran et al., 1989).

Finally, we examine whether these highly-ranked features are associated with the manually-extracted features. We measure the correlation between the features in Table 5.3 and all the manually extracted features. Only those manual features which show a moderate positive or negative correlation with at least one of the automatically extracted features are shown ($r > 0.5$

Feature	PPA-G mean	StrAg mean	*p*
WPM	**53.8**	34.9	0.001
VP proportion	**1.18**	0.94	0.006
aux	**0.058**	0.041	0.009
VP → VBG S	**0.0053**	0.0015	0.0008
light verbs	**0.60**	0.48	0.001
NP → NN NN	0.000085	**0.0023**	0.02
prep on	**0.0071**	0.0026	0.004
PP rate	**0.069**	0.051	0.009
NP → NNP	0.015	**0.024**	0.03

Table 5.3: Group means for those features in Figure 5.3 that are selected more than 50% of the time. Bold font indicates the group with the higher mean. *p* measures the significance of the difference between groups, measured on the complete data set.

		Automatic								
		WPM	VP proportion	aux	VP → VBG S	light verbs	NP → NN NN	prep on	PP rate	NP → NNP
Manual	MLU-word	**0.60**	**0.80**	0.32	0.14	0.21	−0.30	0.14	0.49	−0.46
	MLU-morpheme	**0.59**	**0.79**	0.35	0.13	0.22	−0.32	0.14	0.49	−0.44
	WPM	**0.87**	**0.52**	0.21	0.14	0.30	−0.10	−0.03	0.18	−0.41
	Gram-sentences	0.42	**0.64**	0.34	0.35	0.23	−0.36	0.01	**0.58**	−0.32
	Syn-sentences	0.36	**0.52**	0.31	0.28	0.33	-0.42	−0.10	0.47	−0.14
	Complexity-ratio	0.28	**0.52**	0.24	0.34	−0.03	−0.03	−0.14	0.36	−0.13
	Open-closed-ratio	−0.45	− **0.53**	−0.38	−0.24	−0.44	**0.57**	−0.10	−**0.55**	**0.62**
	Noun-verb-ratio	−0.40	−**0.51**	−0.30	−0.18	−0.40	−0.01	0.01	−0.24	0.48
	Correct-VMI	0.38	**0.62**	0.30	0.30	0.43	−0.33	0.09	**0.50**	−0.21
	Mean-VMI	0.42	**0.61**	**0.75**	0.28	0.37	−0.10	0.10	0.32	−0.35
	Correct-mean-VMI	0.39	**0.58**	**0.72**	0.25	0.38	−0.16	0.14	0.36	−0.15
	Correct-VI	0.26	0.36	0.18	0.24	0.16	−**0.51**	−0.17	0.25	−0.14

Table 5.4: Correlations between the highly ranked, automatically extracted features and the manual features. We only show those manual features for which the absolute value of the correlation with at least one of the automatic features is greater than 0.5. Those correlations with absolute value greater than 0.5 are shown in bold.

or $r < -0.5$).

The only feature in each group which actually attempts to measure the same quantity is speech rate, which is highly correlated. (We also measure noun-verb ratio and mean length of utterance automatically, although they were not selected as highly relevant features. However, the correlations between those variables and the manual ones are 0.93 and 0.98, respectively). Nonetheless, it is interesting to note that many of the automatic features are at least moderately correlated with one or more of the manually extracted features.

For example, speech rate is also positively correlated with the mean length of utterance (both measures of fluency). The VP proportion is correlated with a number of different manual features, and some of these relationships are more intuitive than others. Most sentences are composed of a (relatively short) NP and then a VP, whose length will contribute to both the MLU and the VP proportion. Similarly the VMI increases with elements such as objects and auxiliaries, which also increase the length and number of levels in a VP. More surprising, perhaps, is that VP proportion is also associated with measures of grammaticality, which we make no attempt to evaluate directly (e.g., proportion of grammatical sentences and proportion of verbs with correct VMI).

Use of auxiliary verbs is correlated with VMI, which is to be expected. NP → NN NN is correlated with the open:closed class ratio, and contains only open-class word types. That feature is also negatively correlated with the proportion of correctly inflected verbs. PP rate is positively correlated with the proportion of grammatical sentences and verbs with correct verb morphology index, and negatively correlated with open:closed class ratio, suggesting that narratives with more prepositional phrases also tend to be more grammatically correct. Finally, NP → NNP is correlated with open:closed class ratio, which again makes sense since *Cinderella* is an open-class word, and the alternative (a pronoun) is closed-class.

5.5 Discussion

The issue of whether the agrammatic/nonfluent variant of PPA is qualitatively the same as post-stroke agrammatic/nonfluent aphasia is complicated and somewhat controversial. A major challenge lies in comparing appropriate data sets. Here we have compared the largest two patient groups to date, matched for aphasia severity and education, and using stringent criteria to exclude PPA patients whose primary deficits are not grammatical in nature. Our analysis suggests that there are still observable differences in the narrative language production of the two groups, as measured by both manual features (classification accuracy: 71%) and automatically extracted features (classification accuracy: 76%). Preliminary investigation of the features suggests these differences may lie in the spontaneous production of verb tenses requiring auxiliary verbs and the production of prepositional phrases and pronouns. However, detailed analysis of all the computed features will no doubt enhance this interpretation. A general theme which emerged from this analysis is that the PPA-G participants were less impaired than the StrAg participants (that is, PPA-G participants spoke faster, produced a greater proportion of words in verb phrases, and produced more auxiliaries, prepositional phrases, and pronouns). In future work, it will be informative to compare these results to healthy controls, to determine whether PPA-G speakers fall within the normal range on any of these measures, or whether they are simply *less* impaired than in post-stroke aphasia.

In this chapter we also introduced dependency parse features. Two of these features were selected in the top 10 over 50% of the time, confirming their utility on this task. Future work will involve applying these features to other datasets discussed in this thesis, including PPA and Alzheimer's disease, as well as developing methods to use information from the arguments in the relations.

A major methodological difference in this chapter compared to preceding chapters is the use of manually annotated data. Specifically, dysfluencies and non-narrative utterances were identified by the transcriber and subsequently removed from the analysis. The primary argument against this approach is that we want to ultimately develop a system which *automatically* analyzes speech input for signs of dementia and aphasia, without requiring human intervention. However, my goal in this particular project was primarily to contribute to the scientific question of whether there are syntactic differences between the two patient populations, and in light of that goal it seemed appropriate to use manually annotated data and reduce the probability of parser errors. However, previous results suggest that by throwing away all information related to dysfluency, we may in fact be discarding diagnostically useful information (Fraser et al., 2014b). Future work will involve repeating the analysis with the full, unedited transcripts, as well as comparing features extracted from those transcripts to features extracted from the clean transcripts.

The analysis presented here does not include any direct judgements of grammaticality, which may seem strange, given the task at hand. In more highly structured speech tasks, where there are clear right and wrong answers, a focus on correctness is certainly appropriate. However, in spontaneous (or semi-spontaneous) narrative speech, the number of grammatical errors made may be a function not only of severity but of individual willingness to attempt difficult syntactic structures. Wilson et al. (2010) observed that some of their nonfluent participants never made syntactic errors, apparently *because* they attempted only the simplest syntactic structures. Thus the level of frank agrammatism may vary depending on the compensatory strategies adopted by the individual. Furthermore, the task of reconstructing a hypothetical “target phrase” in a relatively unstructured task such as story-telling becomes very difficult. Given the ungrammatical output *the sister go*, was the target phrase *the sisters go*, or *the sister goes*? The interpretation directly affects the analysis – did the participant omit a plural marker, or a third-person singular inflection? Or perhaps they meant something else entirely; it is impossible to know. Saffran et al. (1989) touch on this issue in their criteria for a system to analyze agrammatic speech, writing, “Minimal assumptions should be made concerning the patients’ intended target utterances, and consequently about the nature of the error produced. Instead, scoring should be based on the structural elements actually present in the utterances. The frequency with which these elements occur in patients’ speech can then be compared to that of a matched normative group." Of course, in some contexts a determiner is obligatory, or a verb is obviously in disagreement with a noun, and in these cases an automated system should be able to capture the error. However, the current approach avoids the very difficult problem of automatic error detection in non-canonical speech while still demonstrating the same (even slightly higher) ability to differentiate between the groups as hand-coded measures of grammatical correctness.

Chapter 6

Speech and language features for the detection of Alzheimer's disease

In previous chapters, we have demonstrated the potential utility of applying natural language processing techniques to the analysis of PPA. Language impairment is the defining characteristic of PPA, and so the rationale for using language samples to classify PPA and its subtypes was clear. In this chapter, we will demonstrate the flexibility of this approach by applying it to the problem of detecting another type of dementia, namely Alzheimer's disease (AD).

AD is more common than PPA, and is in fact the most common type of dementia overall (Alzheimer's Association, 2016). While brain imaging of PPA patients typically reveals a relatively focal degeneration in the language areas of the brain, AD pathology tends to be more global, affecting a wide range of cognitive processes (Kirshner, 2012). The core symptom of AD is memory impairment, typically beginning with an inability to remember new information. However, there are a number of other symptoms associated with the disease, including difficulties with planning or problem-solving, an inability to complete familiar tasks, confusion surrounding time and place, poor judgement, changes in personality, and difficulties with language (Alzheimer's Association, 2016).

In fact, subtle changes in language can be one of the earliest signs of AD (Forbes-McKay and Venneri, 2005; Cuetos et al., 2007; Ahmed et al., 2013). Researchers have suggested that this could be due to the fact that successful discourse production requires a range of cognitive functioning, including planning, attention, and memory (Fleming, 2014). As a result, an analysis of narrative speech can in some cases pick up on changes that are not apparent on a simpler language task, such as confrontation naming (Fleming, 2014).

In this chapter, we present some background information on language in AD and related computational work. We describe the participants and the narrative task — in this case, a picture description task. We then present some new features designed specifically to assess previously reported characteristics of AD speech; namely, repetitions and a reduction in information content. The results of a classification experiment and a factor analysis are discussed in relation to previous work in the field. Finally, we consider the effect of ASR on the features and classification accuracy.

6.1 Background[1]

6.1.1 Language in AD

Although memory impairment is the main symptom of AD, language impairment can be an important marker. Faber-Langendoen et al. (1988) found that 36% of mild AD patients and 100% of severe AD patients had aphasia, according to standard aphasia testing protocols. Ahmed et al. (2013) found that two-thirds of their participants showed subtle, but significant, changes in connected speech production up to a year before their diagnosis of probable AD. Weiner et al. (2008), in a study of 486 AD patients, reported a significant correlation between dementia severity and a number of different linguistic measures, including confrontation naming, articulation, word-finding ability, and semantic fluency.

Declining performance on naming tasks can occur early in the disease progression (Kirsh-

[1] The material presented in Sections 6.1–6.3 was originally published in: Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz (2015). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease* 49(2), pp.407–422.

ner, 2012; Reilly et al., 2011; Kirshner et al., 1984; Taler and Phillips, 2008). Kirshner et al. (1984) found that all the AD participants in their study were impaired on a naming task, even when their language functioning was normal by other measures. Individuals with AD can have difficulty retrieving the names of people and places (Oppenheim, 1994), and may substitute generic terms for more specific ones (Kirshner, 2012; Reilly et al., 2011; Kempler, 1995). Numerous studies have reported a greater impairment in category naming fluency (e.g., naming animals or tools) relative to letter naming fluency (e.g., naming words that start with the letter *R*) (Salmon et al., 1999; Monsch et al., 1992; Adlam et al., 2006), and this finding was supported by a meta-analysis of 153 studies (Henry et al., 2004). There is also some evidence that patients with AD may have more difficulty naming verbs than nouns. Robinson et al. (1996) found that AD participants performed worse on a picture-naming task for verbs than nouns, even when the verbs and nouns were spelled and pronounced the same, and matched for frequency. As a result of word-finding difficulties and a reduction in working vocabulary, the language of AD patients can seem "empty" (Nicholas et al., 1985; Ahmed et al., 2013), and was described by Appell et al. (1982) as "verbose and circuitous, running on with a semblance of fluency, yet incomplete and lacking coherence."

Macro-linguistic language functions, such as understanding metaphor and sarcasm, also tend to deteriorate in AD (Rapp and Wild, 2011). Thematic coherence, or the ability to maintain a theme throughout a discourse, may also be impaired. In a study comparing 9 AD participants to healthy controls and participants with fluent aphasia, Glosser and Deser (1991) found that the AD participants showed a reduction in global coherence in a structured interview task. Blonder et al. (1994) reported a similar result when interviewing 5 AD participants and their healthy spouses.

The effect of AD on syntax is controversial. Some researchers have reported syntactic impairments in AD, while others claim that any apparent deficits are in fact due to difficulties with memory and semantics (Reilly et al., 2011). Several studies have found evidence for a decrease in the syntactic complexity of language in AD (Croisile et al., 1996; Ehrlich et al., 1997; Sajjadi et al., 2012). Croisile et al. (1996) compared oral and written picture descriptions from 22 AD patients and matched controls, and found that the AD patients produced fewer subordinate clauses than controls. Ehrlich et al. (1997) reported a reduced utterance length on narrative tasks administered to 16 AD participants and controls. In a study comparing language production in AD and semantic dementia, Sajjadi et al. (2012) found that their 20 patients with mild AD tended to produce fewer complex syntactic units on both a picture description task and an interview. On the other hand, Kempler et al. (1987) found that 10 individuals with AD used a range of syntactic constructions with the same frequency as control participants in spontaneous conversation, despite showing signs of lexical impairment. Glosser and Deser (1991) similarly did not find any difference in syntactic complexity or correctness between AD patients and controls in spontaneous speech.

There is evidence that language decline in AD is heterogenous. Hodges and Patterson (1995) divided 52 AD patients into three different categories based on dementia severity and assessed their semantic impairment on a number of different tasks. They reported a wide range of performance in the "minimal" and "mild" AD groups. Duong et al. (2005) had 46 AD participants produce narratives based on a single picture and a series of pictures. A cluster analysis subsequently revealed a number of different discourse patterns rather than a single characteristic pattern of impairment. Ahmed et al. (2013) contrasted their findings of heterogenous language decline in connected speech from 15 AD patients with the more predictable patterns of decline seen in primary progressive aphasia.

6.1.2 Related computational work

A relatively small subset of studies on language in AD attempt to quantify the impairments in connected speech using computational techniques. Bucks et al. (2000) conducted a linear discriminant analysis of spontaneous speech from 8 AD participants and 16 healthy controls. They considered eight linguistic features, including part-of-speech (POS) tag frequencies and measures of lexical diversity, and obtained a cross-validation accuracy of 87.5%.

Thomas et al. (2005) classified spontaneous speech samples from 95 AD patients and an unspecified number of controls by treating the problem as an authorship attribution task, and employing a “common N-grams” approach. They were able to distinguish between patients with severe AD and controls with a best accuracy of 94.5%, and between patients with mild AD and controls with a best accuracy of 75.3%. They suggested that closed-class words were particularly informative in their analysis.

Guinn and Habash (2012) built classifiers to distinguish between AD and non-AD language samples using 80 conversations between 31 AD patients and 57 cognitively normal conversation partners. They found that features such as POS tags and measures of lexical diversity were less useful than measuring filled pauses, repetitions, and incomplete words, and achieved a best accuracy of 79.5%.

Meilán et al. (2014) distinguished between 30 AD patients and 36 healthy controls with temporal and acoustic features alone, obtaining an accuracy of 84.8%. For each participant, their speech sample consisted of two sentences read from a screen. The five most discriminating features were percentage of voice breaks, number of voice breaks, number of periods of voice, shimmer, and noise-to-harmonics ratio.

Jarrold et al. (2014) used acoustic features, POS features, and psychologically-motivated word lists to distinguish between semi-structured interview responses from 9 AD participants and 9 controls with an accuracy of 88%. They also confirmed their hypothesis that AD patients would use more pronouns, verbs, and adjectives and fewer nouns than controls.

Rentoumi et al. (2014) considered a slightly different problem: they used computational techniques to differentiate between picture descriptions from AD participants with and without additional vascular pathology ($n = 18$ for each group). They achieved an accuracy of up to 75% when they included frequency unigrams and excluded binary unigrams, syntactic complexity features, measures of vocabulary richness, and information theoretic features.

Orimaye et al. (2014) obtained F-measure scores up to 0.74 using a relatively restricted feature set on transcripts from DementiaBank, although they combined participants with different etiologies, rather than focusing on AD. Prud'hommeaux and Roark (2015) also considered a subset of DementiaBank (130 narratives from the AD group and 130 from controls). Narratives were manually post-processed and limited to 100 words (Prud'hommeaux, 2012). They achieved a best accuracy of 83.2% for distinguishing between the two groups.

Our study differs from previous work in several ways. Along with Orimaye et al. (2014) and Prud'hommeaux and Roark (2015), we consider a much larger sample size than most previous work, giving a more representative sample for machine learning. We also consider a larger number of features to help capture the array of different language impairments that can be seen in AD, and conduct factor analysis to characterize patterns of heterogeneity.

6.2 Data

In this study, we consider narrative speech samples elicited through a picture description task, from participants with probable AD and older, healthy controls. Our data are derived from the DementiaBank corpus, which is part of the larger TalkBank project (MacWhinney et al., 2011). These data were collected between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh. Information about the study cohort is available from Becker et al. (1994). Participants were referred directly from the Benedum Geriatric Center at the University of Pittsburgh Medical Center, and others were recruited through the Allegheny County Medical Society, local neurologists and psychiatrists, and public service messages on local media. To be eligible for inclusion in the study, participants were required to be above 44 years of age, have at least 7 years of education, have no history of nervous system disorders or be taking neuroleptic medication, have an initial Mini-Mental State Exam (MMSE) score of 10 or greater, and be able to give informed consent. Additionally, participants with dementia were required to have a relative or caregiver to act as an informant. All participants received an extensive neuropsychological and physical assessment (see Becker et al. (1994) for complete details). Participants were assigned to the "patient" group primarily based on a history of

	AD ($n = 240$)	Control ($n = 233$)
Age (years)	71.8 (8.5)	65.2 (7.8)
Education (years)	12.5 (2.9)	14.1 (2.4)
Sex (male/female)	82/158	82/151
MMSE	18.5 (5.1)	29.1 (1.1)

Table 6.1: Demographic information for participants with AD and healthy controls. Means and standard deviations are given for each quantity.

cognitive and functional decline, and the results of a mental status examination. In 1992, several years after the study had ended, the final diagnosis of each patient was reviewed on the basis of their clinical record and any additional relevant information (in some cases, autopsy).

From the "Dementia" group, we include participants with a diagnosis of "possible AD" or "probable AD", resulting in 240 samples from 167 participants. We also include control participants, resulting in 233 additional files from 97 speakers. Demographics are given in Table 6.1. We compute averages over individual sessions instead of individual participants in order to capture intra-speaker variation over the five years these data were collected. The two groups are not matched for age and education, which is one limitation of these data.

Narrative speech was elicited using the "Cookie Theft" picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983). This protocol instructs the examiner to show the picture to the patient and say, "Tell me everything you see going on in this picture." The examiner is permitted to encourage the patient to keep going if they do not produce very many words. Each speech sample was recorded, and then manually transcribed at the word level following the TalkBank CHAT (Codes for the Human Analysis of Transcripts) protocol (MacWhinney, 2000). Narratives were segmented into utterances and annotated with filled pauses, paraphasias, and unintelligible words.

From the CHAT transcripts, we keep only the word-level transcription and the utterance segmentation. We discard the morphological analysis, dysfluency annotations, and other associated information, as our goal is to create a fully automated system that does not require the input of a human annotator. Before tagging and parsing the transcripts, we automatically remove short false starts consisting of two letters or fewer (e.g., *The c- cookie jar* would become

The cookie jar) and filled pauses such as *uh*, *um*, *er*, and *ah* (e.g., *The um um boy* would become *The boy*). All other dysfluencies (including repetitions, revisions, paraphasias, and comments about the task) remain in the transcript. The AD participants produce an average of 104.3 (SD: 59.0) words per narrative, while the control participants produce an average of 114.4 (SD: 59.5) words per narrative, although the distribution in both cases is somewhat right-skewed.

Each transcript has an associated audio file, allowing for lexical and acoustic analysis in parallel, which we converted from MP3 to 16-bit mono WAV format with a sampling rate of 16 kHz.

6.3 Detection of AD using manual transcripts

In addition to all the text and acoustic features discussed in Chapter 3, here we also include new features that are designed to capture some of the specific impairments of connected speech in AD that have been noted in the literature. Specifically, we explore techniques for detecting and quantifying repetitive ideas and a reduction in the information content of speech. We also include a new set of acoustic features based on the mel-frequency cepstral coefficients of the audio signal.

6.3.1 New features

Perseverations in speech

One phenomenon that is often noted by the caretakers of AD patients is the occurrence of perseverative behaviour, including in their speech (Bayles et al., 2004; Tomoeda et al., 1996). It has been theorized that this may be a result of an impairment in memory or attention, leading speakers to forget that they had recently uttered the same phrase. Nicholas et al. (1985) found that in a picture description task, AD patients repeated words and phrases more frequently than healthy controls and also more frequently than participants with fluent aphasia. Tomoeda et al. (1996) found also that AD patients were more likely to repeat ideas in a picture description task than healthy controls, and that the frequency of repetitions was not related to severity of dementia.

Recently Barney et al. (2013) proposed an automated technique to detect perseverations in speech samples by modelling the problem as a "motif detection" problem. That is, they transformed the audio signal into discrete subsequences and used existing algorithms to search for repetitive speech patterns. When they tested this method on read speech, they were able to detect repeated sequences with an accuracy of 57% to 71%.

Here, we consider detecting repetitions from the transcripts, rather than from the audio signal. We compute the semantic similarity between utterances by using a bag-of-words model and calculating the cosine distance between each pair of utterances in the transcript. We remove a short list of stopwords, after observing that utterances such as *He is standing on the stool* and *He is holding the cookie* could be considered relatively similar given the common occurrences of *he*, *is*, and *the*. We calculate five features using this information: the average cosine distance between utterances, the minimum cosine distance between utterances, and the number of utterance pairs with cosine distance equal to zero, less than 0.3, and less than 0.5 (normalized by the total number of pairs). Note that if the cosine distance is zero, then the two utterances contain the same words, though not necessarily in the same order.

These features allow us to detect narratives with highly repetitive content, such as in the following example:

all the bad things
sink's overflowing
the stool's going over
and the cookie jar
−guess the little girl she's saying
give me shh
and the sink's overflowing
I might not be very observant but I don't see anything else

Information units

Picture description tasks can be scored in different ways. In the BDAE, examiners count seven different variables, mostly related to syntax: the total number of utterances, the number of empty utterances, subclausal utterances, single clause utterances, multi-clause utterances, agrammatic deletions, and a complexity index (number of clauses per utterance) (Goodglass and Kaplan, 1983). Many of these concepts are related to the syntactic features we already measure. However, in the Western Aphasia Battery (WAB), a similar picture description task is judged on two dimensions: fluency and information content (Kertesz, 1982). Numerous studies have found that individuals with probable AD tend to produce narratives with lower information content than healthy controls (Giles et al., 1996; Croisile et al., 1996; Forbes-McKay and Venneri, 2005; Ahmed et al., 2013).

Giles et al. (1996) defined an *information unit* as “the smallest non-redundant fact or inference about the picture, which may range in size from e.g., plural /s/ to a phrase.” They did not define any criteria for whether the information conveyed was relevant to the picture itself. In contrast, Croisile et al. (1996) defined a set of specific information units that they expected to be conveyed, based on a survey of earlier literature. These information units are listed in Table 6.2. They found a significant difference between the number of information units produced by controls and by AD participants in both verbal and written picture descriptions. Ahmed et al. (2013) used the same list of information units, as well as *idea density*, which was defined as the number of information units divided by the the total number of words in the sample, and *efficiency*, defined as the number of information units divided by the total sample time in seconds. They found a significant downward trend on a composite score consisting of all the semantic content features when they computed it for participants with mild cognitive impairment, mild AD, and moderate AD.

In the studies mentioned above, information units were identified and counted manually. There has also been some progress towards the automatic scoring of picture descriptions. Hakkani-Tür et al. (2010) compared the scores they obtained through automatic speech recog-

Subjects: BOY, GIRL, WOMAN
Places: KITCHEN, EXTERIOR SEEN THROUGH WINDOW
Objects: COOKIE, JAR, STOOL, SINK, PLATE, DISHCLOTH, WATER, WINDOW, CUPBOARD, DISHES, CURTAINS
Actions: BOY TAKING OR STEALING, BOY OR STOOL FALLING, WOMAN DRYING OR WASHING DISHES/PLATE, WATER OVERFLOWING OR SPILLING, ACTION PERFORMED BY THE GIRL, WOMAN UNCONCERNED BY THE OVERFLOWING, WOMAN INDIFFERENT TO THE CHILDREN

Table 6.2: Information units for the scoring of information content in the Cookie Theft picture (from Croisile et al. (1996)).

nition and an automated scoring system to manual scores, and found very high correlation between the automatically computed scores on ASR and reference transcripts, and reasonably high correlation between those scores and the manual scores. To compute the scores, they measured the unigram recall between the picture descriptions and a list of pre-defined information units. However, it is not clear how they handled multi-word phrases ("on the beach" is one example from their task), and they did not take into account any synonyms, hypernyms, and so on. In other related work, Pakhomov et al. (2010b) calculated information units in Cookie Theft picture descriptions from patients with frontotemporal dementia by matching sequences of 1 to 4 words with a list of pre-defined words and phrases. They did not find a significant difference between their patient groups on this measure.

There has also been related work on the automated scoring of story retelling tasks for detecting MCI (Prud'hommeaux and Roark, 2011, 2012; Lehr et al., 2012). The general idea behind this method is to identify story elements in the participant narratives by using forced alignment with the source narrative (i.e., the original story). The authors achieve high recall and precision on the extraction of story elements, and their diagnostic accuracies are similar to those using the manually extracted scores. However, this method is less suitable for determining the information content of picture descriptions, because there is no "source narrative" in the picture description task, and therefore, the exact lexical content is not defined. Prud'hommeaux and Roark (2015) show that one way to resolve this issue is to choose one of the control narratives as the "gold standard". The benefit of their approach is that it is data-driven: it does not require an expert annotator to compile a list of expected information units. The disadvantage is that it is dependent on the content of the specific data set, and the criteria for selecting the gold standard transcript.

Subjects, places, and objects The approach that we take here is to search for words that are related to each of the expected information units. We start by considering the subject, place, and object concepts in Table 6.2. Since these concepts all correspond to nouns, it is relatively easy to construct a set of words that we might expect the participants to use when referencing these concepts. Starting with each word given in the table, we first constructed a set of related words by looking at the WordNet synset for the most relevant sense of the word (Fellbaum, 1998). We then added words from the direct hypernym and hyponym synsets as well. Finally we manually reviewed the sets to remove words that were unlikely to be used in the context of the picture (e.g., *binary compound* for WATER) and to add words which were not directly related in WordNet but which seemed likely given the picture (e.g., *mother* for WOMAN). The final set of words relating to each subject, place, and object information unit is given in Table 6.3.

Obviously, these lists of words do not cover the entire set of possible words that could be used to describe each concept, which is one limitation of this method. Another difficulty we encountered was potential overlap between concepts. For example, the word *female* could be used to describe both the GIRL and the WOMAN, and *dish* could refer to either the PLATE in the woman's hand, or the DISHES on the counter. Since our word-counting method is too simplistic to take context into account, we simply put each word with the concept it is most likely to refer to, with the understanding that this may be a source of error.

Given the sets of words in Table 6.3, we calculate two types of features. The first is a binary information unit for each concept. For example, if the speaker uses the word *boy*, *son*, or *brother*, then the value of the information unit for the concept BOY is one, otherwise it is

Classes	Concepts	Words
Subjects	BOY	*boy, son, brother*
	GIRL	*girl, daughter, sister*
	WOMAN	*woman, female, adult, grownup, lady, mother*
Places	KITCHEN	*kitchen, room*
	EXTERIOR	*outside, outdoors, yard, backyard, garden, driveway, path, tree, bush*
Objects	COOKIE	*cookie, biscuit, cake, treat*
	JAR	*jar, container, crock, pot*
	STOOL	*stool, seat, chair, ladder*
	SINK	*sink, basin, washbasin, washbowl, washstand, tap*
	PLATE	*plate*
	DISHCLOTH	*dishcloth, dishrag, towel, rag, cloth*
	WATER	*water, dishwater, liquid*
	WINDOW	*window, frame, glass*
	CUPBOARD	*cupboard, closet, shelf*
	DISH(ES)	*dish, cup, counter*
	CURTAIN(S)	*curtain, drape, drapery, blind, screen*

Table 6.3: Keywords corresponding to each of the concepts (i.e. subjects, places, and objects) in Table 6.2.

zero. We have 16 information unit features, corresponding to the 16 concepts.

However, it could also be informative to know specifically which of the words was used to describe a concept. For example, someone with word-finding difficulties might say *room* instead of *kitchen*, or *container* instead of *jar*. To capture this information, we also compute an individual frequency count for each of the words in the third column of Table 6.3. We call these features "key words", and they are integer-valued frequency counts. Previous work has shown the utility of simple binary and frequency unigrams (Rentoumi et al., 2014; Garrard et al., 2014). Rather than considering the space of all possible unigrams, we have considered only this smaller set which we have deemed to be relevant to the expected information content, to avoid problems of data sparsity, and to help improve the interpretability of the selected features.

Actions The information units describing actions are more of a challenge to encode. We begin with a fairly simple approach that could be easily extended to account for more complex lexical variation in future work. We extract the typed dependencies of each utterance using the Stanford parser, and attempt to match them with each of the actions listed in Table 6.2. Table 6.4 shows the mapping from concept to dependency. Note that in the Stanford parser, *nsubj* represents the subject of a verb. We consider the verb lemma in each case, to allow for variations such as *The boy falls off the stool* and *The boy is falling off the stool*. In the case of the boy or the woman, we also allow the substitution of the pronouns *he* or *she*, respectively. However, in the case of the girl, we do not allow the substitution of a pronoun. Since the action is unspecified, allowing a pronoun in this case could create matches with the woman's actions as well. Similarly, we do not allow the stool to be replaced by *it*, because the rule becomes too general (for example, it could match with *The water is running and it is falling all over the floor*).

On the other hand, we underspecify the action of the woman washing the dishes/plate, allowing any utterance that describes the woman washing or drying any object, since it seems likely that in most cases this would be an appropriate match and we can reduce the risk of false negatives by keeping the rule more general. We also do not specify the subject of *overflow* and *spill*, since in most cases the presence of the verb alone indicates the appropriate information unit, regardless of whether the speaker chooses to say *the water is overflowing*, *the sink is overflowing*, *it is overflowing*, etc. In the current work, we do not include the two information units corresponding to the woman's indifference to or ignorance of the water overflowing or the children stealing the cookies.

False negatives are an issue with this approach, for a number of reasons. As with the subjects, objects, and places, there are a number of ways to represent the same concepts, except here the problem is twice as large since each action has at least a subject and a verb. For example, the boy could be described as *the boy*, *he*, *the kid*, etc., and the action of taking the cookie could be described as *taking*, *stealing*, *getting*, *grabbing*, etc. Then these two concepts can be paired in any combination. Furthermore, depending on how the action is expressed, the parser may not correctly identify the verb and its object. One simple example is the sentence *The boy is going to fall*, which produces the dependency nsubj(*go, boy*) rather than nsubj(*fall,*

Actions	Dependencies
BOY TAKING OR STEALING	nsubj(*take, boy*) or nsubj(*take, he*) or nsubj(*steal, boy*) or nsubj(*steal, he*)
BOY OR STOOL FALLING	nsubj(*fall, boy*) or nsubj(*fall, he*) or nsubj(*fall, stool*)
WOMAN DRYING OR WASHING DISHES/PLATE	nsubj(*dry, woman*) or nsubj(*dry, she*) or nsubj(*wash, woman*) or nsubj(*wash, she*)
WATER OVERFLOWING OR SPILLING	nsubj(*overflow,* *) or nsubj(*spill,* *)
ACTION PERFORMED BY THE GIRL	nsubj(*, *girl*)
WOMAN UNCONCERNED BY THE OVERFLOWING	–
WOMAN INDIFFERENT TO THE CHILDREN	–

Table 6.4: Dependencies corresponding to each of the actions in Table 6.2.

boy).

To validate this approach, a certified speech-language pathologist also annotated these information units over a random 5% of the data and we compared these against the automatically identified units. There was an observed agreement of 98.02%, which corresponds to a Cohen's kappa coefficient of 0.804.

MFCC features

We also consider a set of additional acoustic features, based on the mel-frequency cepstral coefficients (MFCCs). MFCCs are features that are commonly used in speech recognition. In a nutshell, the *spectrum* of an audio signal is its representation in the frequency domain, and is obtained by taking a Fourier transform of the time domain signal. The *cepstrum* of the signal is obtained by then transforming the logarithm of the spectrum using a discrete cosine transform (Quatieri, 2002). This separates the 'source' of a signal (i.e., the energy of the lungs) from the 'filter' of the signal (i.e., the upper vocal tract, in which phonemes are differentiated phonologically). The mel-scale simply refers to the fact that the frequencies are windowed based on a perceptual scale rather than a linear scale. As frequency increases, the human ear is less sensitive to differences in frequency, and so the intervals on the mel-scale increase

(Jurafsky and Martin, 2000).

The first several MFCCs represent information about the vocal tract filter and are often used as part of the input to a speech recognition system, although they have also been used as features in classification problems (e.g., Cummins et al. (2011)). We consider the mean, variance, skewness, and kurtosis of the energy and the first 13 MFCCss through time, as well as the mean, variance, skewness, and kurtosis of the first and second derivatives of the energy and MFCCs. We also calculate the kurtosis and skewness of the means across MFCCs.

6.3.2 Methods

We train a logistic regression classifier to distinguish between AD patients and healthy controls. We perform a 10-fold cross-validation procedure in which a unique 10% of the data are used in each iteration for evaluation, and the remaining 90% are used to select the most useful features (from the 370 available) and construct our models. The reported accuracy is an average across the 10 folds. In a given fold, data from any individual speaker can occur in the test set or the training set, but not both. In order to optimize the ratio of training examples to their dimensionality, we select the N features with the highest Pearson's correlation coefficients between each feature and the binary class.

6.3.3 Results

Figure 6.1 shows the average accuracies (and standard deviations) for the logistic regression method. The maximum average accuracy (81.92%) in distinguishing between AD and controls is achieved with the 35 top-ranked features. The accuracy remains relatively constant until we choose a feature set of size 50 (accuracy = 78.72%), after which it drops off sharply. As a result, we use the top 50 features in our factor analysis. Those features, and their correlation with diagnosis, are shown in Table 6.5. Using all 370 features, the logistic regression method obtains 58.51% accuracy on average, which reinforces the need to do feature selection given a high-dimensional feature space such as this.

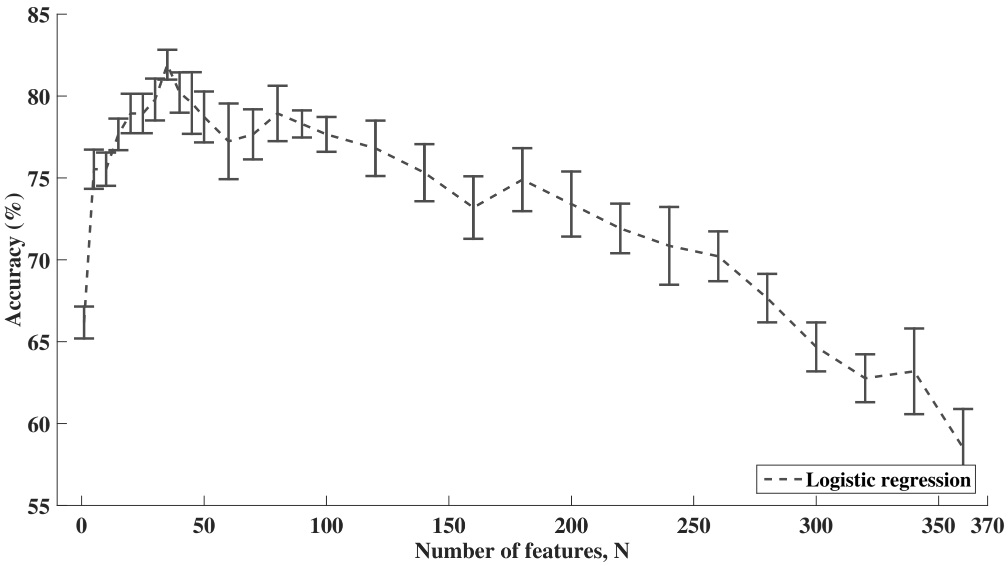


Figure 6.1: Classification results for AD versus healthy controls.

6.3.4 Factor analysis

To help discover the underlying structure of the data, we conduct an exploratory factor analysis. Since our data do not satisfy the assumption of multivariate normality, we use the method of principal axis factors (PAF), as recommended by Fabrigar et al. (1999) . We include 50 features in the factor analysis, as discussed in the previous section.

A scree test suggests that four factors are sufficient to account for the majority of the variance. To interpret the factor structure, it is customary to perform a rotation. Although varimax is the most popular rotation algorithm, it is an orthogonal rotation and is therefore guaranteed to produce uncorrelated factors. To fairly examine the degree of heterogeneity of linguistic impairments in patients with AD, we chose promax, an oblique rotation which allows factors to be correlated with each other (Costello and Osborne, 2005).

6.3.5 Factor loadings

Feature loadings on the four factors are presented in Table 6.5. Factor signs were deliberately set such that higher factor scores reflect greater impairment. As is customary in exploratory

Feature	*R*	Factor1	Factor2	Factor3	Factor4
Pronoun:noun ratio	0.35	**1.01**		**−0.32**	
NP → PRP	0.37	**0.88**		−0.24	
Frequency	0.34	**0.74**			
Adverbs	0.31	**0.51**			0.19
ADVP → RB	0.30	**0.44**		0.10	
Verb frequency	0.21	**0.39**		0.13	
Nouns	−0.27	**−0.97**		**0.37**	
Word length	−0.41	**−0.60**		−0.13	
NP → DT NN	0.10	**−0.52**			−0.19
Honoré's statistic	−0.25	**−0.46**	−0.14	−0.14	**0.33**
Inflected verbs	−0.19	**−0.39**			
Average cosine distance	−0.19	**−0.33**		−0.15	0.13
Skewness(MFCC 1)	0.22		**0.95**		
Skewness(MFCC 2)	0.20		**0.87**		−0.14
Kurtosis(MFCC 5)	0.19		**0.78**		
Kurtosis(VEL(MFCC 3))	0.24	−0.17	**0.44**		0.24
Phonation rate	−0.21	0.16	**−0.62**		−0.28
Skewness(MFCC 8)	−0.22		**−0.39**		−0.13
Not-in-dictionary	0.38	−0.14		**0.53**	0.26
ROOT → FRAG	0.23	−0.15		**0.36**	0.19
Verbs	−0.29	**0.38**		**−1.05**	0.20
VP rate	−0.19	**0.37**		**−0.95**	**0.32**
VP → AUX VP	−0.23	−0.16		**−0.56**	0.18
VP → VBG	−0.27	−0.28		**−0.34**	0.21
Key word: *window*	−0.29			0.20	**−0.79**
Info unit: WINDOW	−0.32			0.12	**−0.63**
Key word: *sink*	−0.23				**−0.62**
Key word: *cookie*	−0.23		0.13		**−0.61**
PP proportion	−0.21			0.18	**−0.61**
Key word: *curtain*	−0.25				**−0.56**
PP rate	−0.21			0.19	**−0.55**
Info unit: CURTAIN	−0.26				**−0.53**
Key word: *counter*	−0.18			0.14	**−0.47**
Info unit: COOKIE	−0.24				**−0.46**
Info unit: SINK	−0.31				**−0.43**
Info unit: GIRL	−0.30				**−0.42**
Info unit: GIRL'S ACTION	−0.25		0.13	−0.12	**−0.36**
Info unit: DISH	−0.24	−0.12			−0.29
Key word: *stool*	−0.28	−0.15			−0.29
Key word: *mother*	−0.32			−0.27	−0.26
Info unit: STOOL	−0.32	−0.29			−0.21
Skewness(MFCC 12)	−0.19				−0.18
Info unit: WOMAN	−0.29			−0.16	−0.18
VP → VBG PP	−0.34	−0.19		**−0.30**	−0.12
VP → IN S	−0.20				−0.10
VP → AUX ADJP	−0.19	−0.11			
VP → AUX	0.20	0.28			
VP → VBD NP	0.19				
Cosine cutoff: 0.5	0.19		0.15	0.14	
INTJ → UH	0.18	0.25			

Table 6.5: Correlations with diagnosis (first column) and promax factor loadings. Loadings less than 0.1 are excluded. Bold font indicates a loading with an absolute value greater than 0.3.

factor analysis, we name and present a subjective interpretation of the factors, below.

Factor 1: Semantic impairment

All of the high loadings reflect characteristics of semantically impoverished language, similar to that seen in the semantic variant of PPA (Gorno-Tempini et al., 2004). Individuals scoring high on this factor produce many pronouns (+NP → PRP, +pronoun ratio) and few nouns (−nouns), and are biased towards shorter (−word length) and higher frequency words (+frequency, +verb frequency). They also use a less diverse vocabulary (−Honoré's statistic) and exhibit increased repetition of content (−cosine distance).

Pronouns and high frequency words suggest empty, vague, or non-specific speech. A decrease in the proportion of nouns and an increase in the proportion of verbs is the same pattern as seen in sv-PPA (Harciarek and Kertesz, 2011; Wilson et al., 2010). Individuals with a semantic impairment may have difficulty accessing more specific nouns and verbs, and as a result may replace them with generic, high-frequency substitutes (e.g., Meteyard et al. (2014)). Negative Honoré's statistic suggests low lexical diversity, which has been observed in anomic aphasia (Fergadiotis and Wright, 2011), and negative cosine distance suggests high repetition, which bears similarity to the "stereotypic thematic perseverations" seen in sv-PPA (Harciarek and Kertesz, 2011). Examples of the adverbial construction (+ADVP → RB) include *the little girl's reaching up there* and *a tree coming up here*; that is, the adverb serves a deictic purpose, which is more common amongst aphasics with a semantic impairment (Varley, 1993).

Factor 2: Acoustic abnormality

All high loadings here relate to features derived from acoustic analysis. All but one of these refer to either the skewness or kurtosis of individual Mel-frequency cepstral coefficients, whose perceptual values may not be distinguishable to humans, and whose anatomical basis depends on interpreting the vocal tract within a source-filter model of speech production. The remaining feature, phonation rate, is the proportion of a narrative that is vocalized; low values here refer to more time being spent silently, as in a pause.

Factor 3: Syntactic impairment

This factor appears to reflect a syntactic deficit somewhat reminiscent of such conditions as Broca's aphasia and nonfluent/agrammatic PPA. High-scoring patients produced fewer verbs, which is typical of agrammatic patients (e.g., Saffran et al. (1989), and Thompson et al. (1997a)). They also produced fewer auxiliary verbs, and fewer gerunds and participles. Patients with Broca's aphasia often omit auxiliaries and use only the simplest verb tenses (e.g., *he reach* might be preferred over *he is reaching*, which requires an auxiliary and a participle) (Bastiaanse and Thompson, 2003; Menn, 1995). Additionally, they produced more sentence fragments, and more words tagged as "not in dictionary," which could include phonological paraphasias, distortions, and unrecognizable words (note that the automatic analysis cannot distinguish between these different language phenomena). We note that while these deficits resemble Broca's aphasia and nfv-PPA in their form, they are less severe, seldom reaching the point of frank agrammatism or "telegraphic" speech seen in those disorders. Presumably this reflects the fact that cortical damage to language centres in Alzheimer's disease is less severe than in those conditions.

Factor 4: Information impairment

This factor primarily includes mention of key words and information units. Patients with high scores produced relatively uninformative picture descriptions, failing to mention key concepts. This factor differs from Factor 1 in that the relevant features do not describe generic properties of the words, such as their frequency or part-of-speech, but rather their appropriateness and specific semantic relevance to the task at hand. These participants also lacked prepositional phrases, which reflects a lower level of detail in their descriptions. There may be some relation between the absence of certain information units and a reduction in prepositional phrases. For example, both the information unit SINK and key word *sink* are negatively weighted on this factor. When we examined the control transcripts, we found that in 57% of cases, the word *sink* appeared as the object of a prepositional phrase (e.g., *water's overflowing in the sink*, *the water is spilling out from the sink*, *the mother's working at the sink*). This potential connection between the omission of certain content words and a reduction in prepositional phrases will require further investigation in future work.

6.3.6 Relationships among factors

Figure 6.2 shows pairwise scatterplots of individual transcript scores on each factor. All four factors are significantly different between groups, which is not surprising given that the features comprising the factors were pre-selected for their association with diagnosis. Of greater interest is the degree of correlation among the factors, which can be estimated by the oblique promax rotation, as opposed to orthogonal rotations such as varimax that guarantee uncorrelated factors. The correlation coefficients among all samples, and limited to AD and control subjects alone, are given in each plot.

An intriguing result is the correlation between Factor 1 (semantic impairment) and Factor 3 (syntactic impairment). These factors are moderately correlated in the control group ($R = 0.42$) but much less correlated in the AD group ($R = 0.19$). This suggests that in cognitively normal individuals, semantic and syntactic abilities are somewhat linked, but when these abilities decline in AD, there can be an asymmetry to the impairment. This may be attributable to damage specific to networks responsible for distinct aspects of linguistic competence (see Section 6.3.7).

Another pattern is seen when we consider Factor 4 (information impairment). Factor 1 and Factor 4 are more highly correlated in the AD group ($R = 0.49$) than the control group (0.18). Similarly, Factor 3 and Factor 4 are more highly correlated in the AD group ($R = 0.31$) than the control group ($R = -0.097$). Since information is expressed through both syntax and semantics, we hypothesize that difficulty in either Factor 1 or Factor 3 would also imply impairment in conveying information. That a similar correlation is not seen in the control data

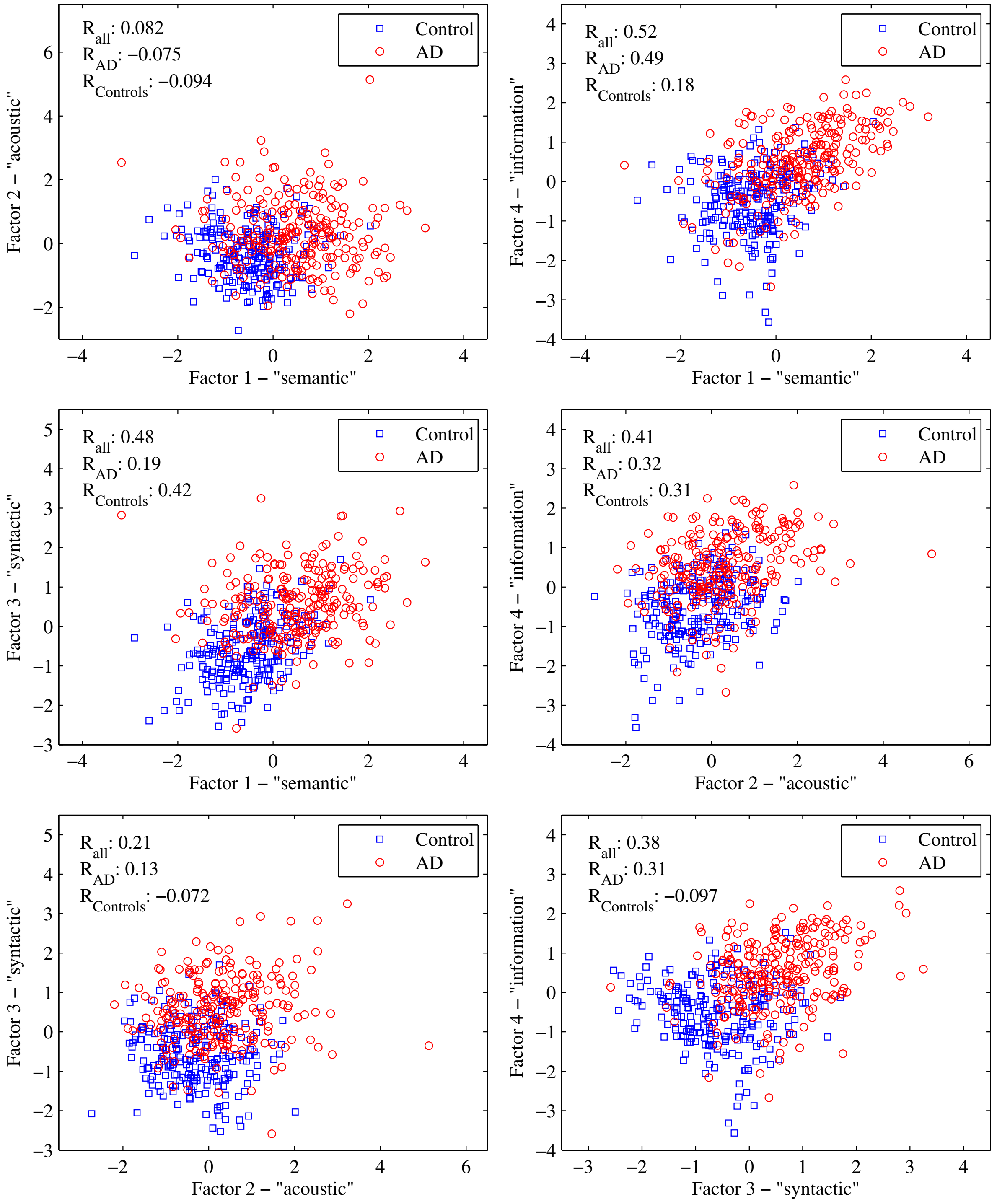


Figure 6.2: Pair-wise combinations of the four factors. Correlation coefficients are given for the entire data set, as well as just AD and control groups.

might be due to the fact that the task is generally easy for controls, and our method for scoring information units is very simplistic. For example, when we look at individual transcripts, we find a control participant who uses a number of pronouns and generic, high-frequency words (e.g., *thing*, *something*). However, this semantic "impairment" does not prevent the participant from mentioning most of the relevant information units. Conversely, a control participant who uses more detailed language will not be able to increase their information score beyond what our simple keyword-spotting algorithm can detect.

Factor 2 (acoustic abnormality) is uncorrelated with Factor 1 (semantic impairment) and Factor 3 (syntactic impairment) in both the AD and control groups. Factor 2 is moderately correlated with Factor 4 (information impairment) in the AD ($R = 0.32$) and control ($R = 0.31$) groups. This would suggest that participants who spoke slower (phonation rate) or had some other acoustic irregularity (MFCC features) also offered less information, regardless of whether they were a patient or a control.

6.3.7 Discussion

Although language impairment is a secondary cognitive symptom of AD, numerous studies have shown that language skills become abnormal relatively early in AD, and can serve as a sensitive index of disease severity over time. The present study employs a combination of automated quantification of language with a modern machine-learning classification approach to accurately distinguish between healthy controls and patients with Alzheimer's disease on the basis of short speech samples from a picture description task.

The relatively short length (roughly 100 words) of the picture descriptions is one drawback of the current study. Other researchers have suggested that 150 words is an acceptable minimum length for language analysis (Sajjadi et al., 2012; Saffran et al., 1989). Our classification results indicate that there is still valuable information to be found in the short samples, confirming previous studies using picture descriptions to assess language in dementia (Rentoumi et al., 2014; Garrard et al., 2014; Wilson et al., 2010); however, we expect that the accuracies of each feature value would increase as the length of the sample increased. We also emphasize that our findings here do not necessarily generalize to other spoken language tasks. For example, Sajjadi et al. (2012) found that the picture description task was more sensitive to semantic impairment in AD, while an interview format was more sensitive to syntactic measures.

Factor analysis reveals that our relatively large set of linguistic measures can be mapped to 4 latent variables, broadly representing syntax/fluency, semantics, acoustic differences, and other information content. Each of these language domains has been reported separately to be altered in patients with AD and MCI (Meilán et al., 2014; Snowdon et al., 1996; de Lira et al., 2011; Kemper et al., 2001), but the relationships between separate domains of impairment have seldom been characterized. Many previous studies have relied upon labour-intensive manual analyses of language samples, emphasizing particular aspects according to the research interests of the authors. The large heterogeneity in language impairments reported across studies leaves open the question of whether there is a single quantifiable aspect of spoken language output that is particularly diagnostic of Alzheimer's disease. The heterogeneity of reported impairments is presumably attributable to two sources of variability: differences between individual patients, and differences in the methods and hypotheses employed by the authors of the studies.

For the present study, we aimed to capture as broad a spectrum of linguistic variables as possible using fully automated analysis of transcripts and acoustic recordings, from a relatively large sample of picture descriptions. This approach, while potentially missing some useful variables that require manual analysis, can characterize the level of heterogeneity present across individual patients tested with a consistent protocol. Previous studies of cognitive decline in MCI/AD have highlighted considerable heterogeneity among patients, but have tended to view language as a fairly unitary construct along with other cognitive domains including episodic memory and visuospatial cognition. Despite this heterogeneity, there seems to be a typical pattern of decline in AD particularly driven by impairments in episodic memory (Lambon Ralph et al., 2003b), with other domains affected in a more limited set of patients.

Variability in the cognitive presentation of AD is to be expected, given that different patients have damage to different parts of the cortex. To date, little work has been done to characterize the relationship between impairments on specific subdomains of language and cortical atrophy in AD. This situation contrasts strongly with that of primary progressive aphasia (PPA). As discussed in Chapter 3, a current consensus among researchers holds that PPA can be clearly divided into at least three subtypes characterized by impairments to distinct aspects of language (Gorno-Tempini et al., 2004). The double dissociation of syntactic and semantic impairments in PPA highlights the fact that distinct brain networks make unique contributions to linguistic competence. Thus, even though the episodic memory impairment dominates the cognitive profile of AD patients, variability in cortical involvement across patients should differentially impact the same subdomains of language that are affected in PPA. However, this variability in AD is likely to be more subtle than between PPA subtypes, as the patients do not fall into clearly distinguishable diagnostic categories.

Although memory impairment may be the definitive symptom for the diagnosis of AD, it is not necessarily the most sensitive index of cognitive function and response to intervention. Language function especially degrades as the disease progresses through moderate and severe stages (Cummings et al., 1985; Feldman and Woodward, 2005), and has been shown to improve in response to successful treatment with acetylcholinesterase inhibitors (Ferris and Farlow, 2013). Therefore, computational analyses of naturalistic language may ultimately provide a means to monitor changes in cognitive status over the course of the disease, as well as responsiveness to interventions, and can thus serve as a useful clinical tool for purposes well beyond diagnosis.

6.4 Detection of AD using ASR transcripts

In this section, we consider using ASR transcripts (and no manually-derived information) to classify AD versus healthy controls. One benefit to working with the larger DementiaBank corpus is that it contains enough data to train a limited-vocabulary automatic speech recognition (ASR) system, rather than having to adapt an existing system to our data.

The speech recognition experiments were carried out by a collaborator, Luke Zhou, and are described in Zhou et al. (2016). Briefly, he trained DementiaBank-specific ASR models using the open-source speech recognition toolkit Kaldi (Povey et al., 2011). The corpus was divided into 10 folds, and in each fold the ASR models were trained using the test set (i.e., audio files and transcripts from 90% of the entire corpus) and then used to generate transcripts for the remaining 10% of the corpus. This process was repeated until all the DementiaBank samples had been recognized. In the classification experiments which follow, we use the same data partitions as were used in the ASR experiments.

In the ASR optimization, various combinations of model complexity, language model weight, and insertion penalty were explored (see Zhou et al. (2016) for details). In this section, we use the transcripts from the most successful recognition experiment (corresponding to the *tri3* training model, language model weight = 20, and insertion penalty = 0.0). These transcripts have an average word error rate (WER) of 36.6%. This is a substantial improvement over the WERs reported in Chapter 4, on the PPA data set, and is commensurate with the previous work discussed in Section 2.6. Furthermore, the recognition procedure used by Zhou et al. (2016) operated on each utterance separately, rather than on the sample as a whole, so utterance segmentation is not required.[2] Thus, we expect that more of the resultant text features may be relevant in these ASR transcripts, compared to our previous results.

6.4.1 Features

To explore how the discriminative properties of different features are affected in the ASR transcripts, we construct the plots seen in Figure 6.3 and Figure 6.4.[3] Each bubble represents a

[2] Note that this is in some sense a limitation of the method, since a fully automated system would presumably analyze the speech sample as a whole. However, this procedure was necessary due to a technical requirement in Kaldi for transcripts to contain no more than 10-20 words.

[3] Interactive, zoomable versions of these plots are available at `http://www.cs.toronto.edu/~kfraser/thesis_bubbles.html`

feature. The size and colour of the bubble represents the correlation between the feature values in the manual and ASR transcripts (larger radius and warmer colour indicate higher correlation. No feature values were negatively correlated in the two sets of transcripts.). The axes represent the degree of significance between the AD and control groups in the manual and ASR transcripts. For perfectly recognized speech, we would expect that the bubbles would line up along $y = x$, with each pair of features being perfectly correlated. Obviously, this is not the case here. Instead, bubbles in the upper-right quadrant of Figure 6.3 are features which have no diagnostic utility in either the manual or ASR transcripts (p close to 1). Bubbles in the lower-right correspond to features that are more useful (lower p) in the ASR transcripts than the manual transcripts. Bubbles in the upper-left are more useful in the manual transcripts than the ASR transcripts, and bubbles in the lower-left (as p approaches 0 on each axis) are features that are useful in both sets of transcripts.

In Figure 6.3, most visible bubbles are *not* significant at $\alpha = 0.05$. However, when we use a log-scale axis as in Figure 6.4, features which differ significantly between the groups become visible. In Figure 6.4, we observe fewer examples of features which have very low correlation (small, dark blue bubbles) in the ASR and manual transcripts; however, we also observe fewer features which are very highly correlated (large, red bubbles). The bubbles tend to follow a linear trend, that is, there is a tendency for highly relevant features in the manual transcripts to also be highly relevant in the ASR transcripts. However, most of the features lie above $y = x$, indicating that, in general, features are less significant in the ASR transcripts.

The location and size of the feature bubbles in the plot can help us predict how the set of selected features in the ASR transcripts will change relative to the manual transcripts. For example, we can see that NID, or ‘not-in-dictionary’ words (the small blue bubble in the upper-left quadrant of Figure 6.4), is highly significant between the groups in the manual transcripts, but much less so in the ASR transcripts. This makes sense, considering that any particular NID word in the test data is unlikely to have appeared in the training data, and is likely to be misrecognized as a real (i.e., more probable) word. We can also see that the correlation for this feature is low across the two sets of transcripts, suggesting that the feature is not identifying the same NID words in the ASR and manual transcripts. In contrast, word frequency and VP → VBG PP are highly significant between the groups in both the manual and ASR transcripts, and so we expect those features to be selected in both cases.

There is no obvious pattern to the features which are significant in the ASR transcripts – they include measures of frequency, information content, grammatical productions, and POS tags. This is in contrast to our findings with the very noisy PPA transcripts in Chapter 4, where only the psycholinguistic features were still relevant in the ASR transcripts. This may be partly due to the much lower WER here.

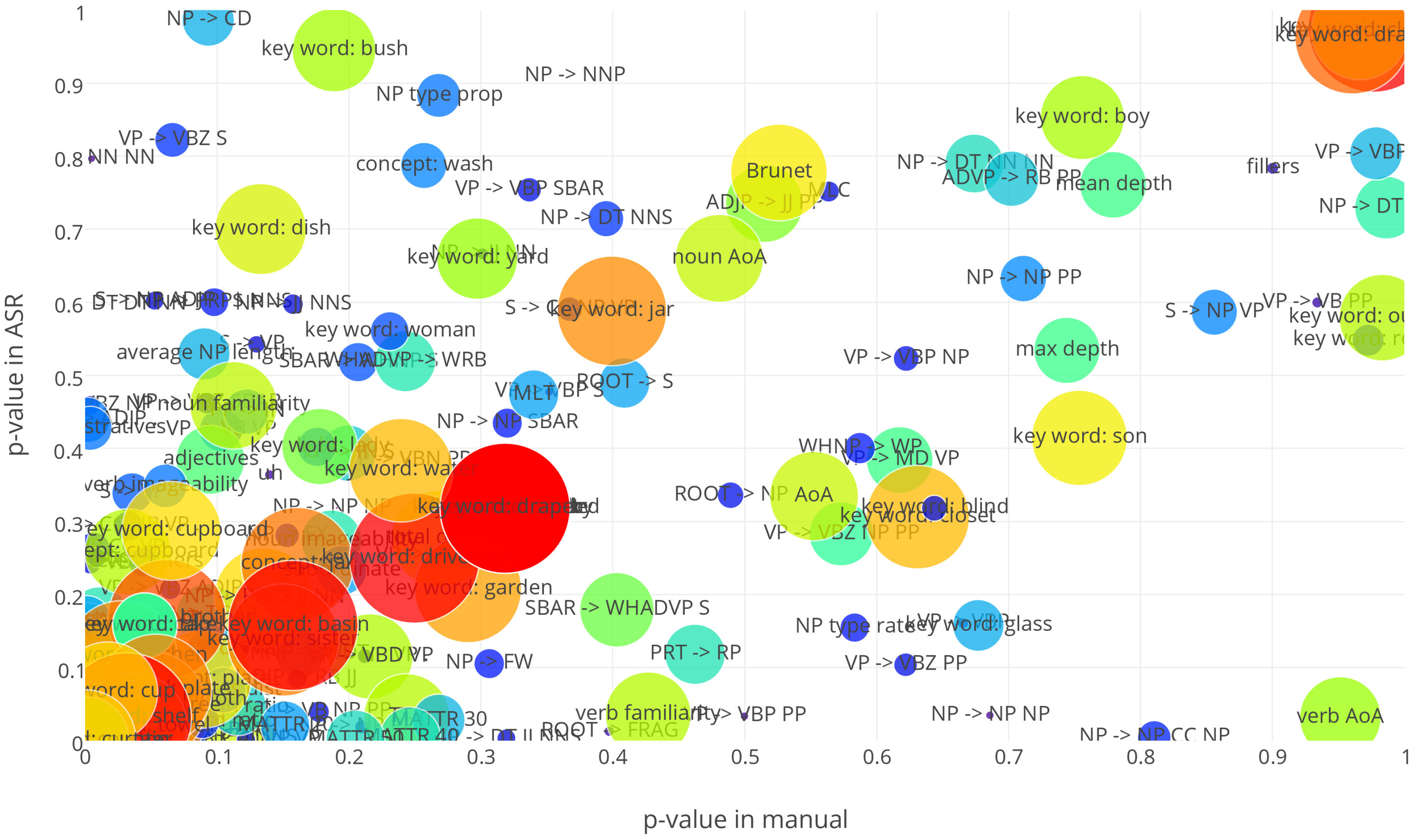


Figure 6.3: Bubbles represent features; large red bubbles indicate features that are highly correlated in the ASR and manual transcripts, while small blue bubbles indicate low correlation. Linearly scaled axes show features which are not significant between the control and AD groups (p close to 1).

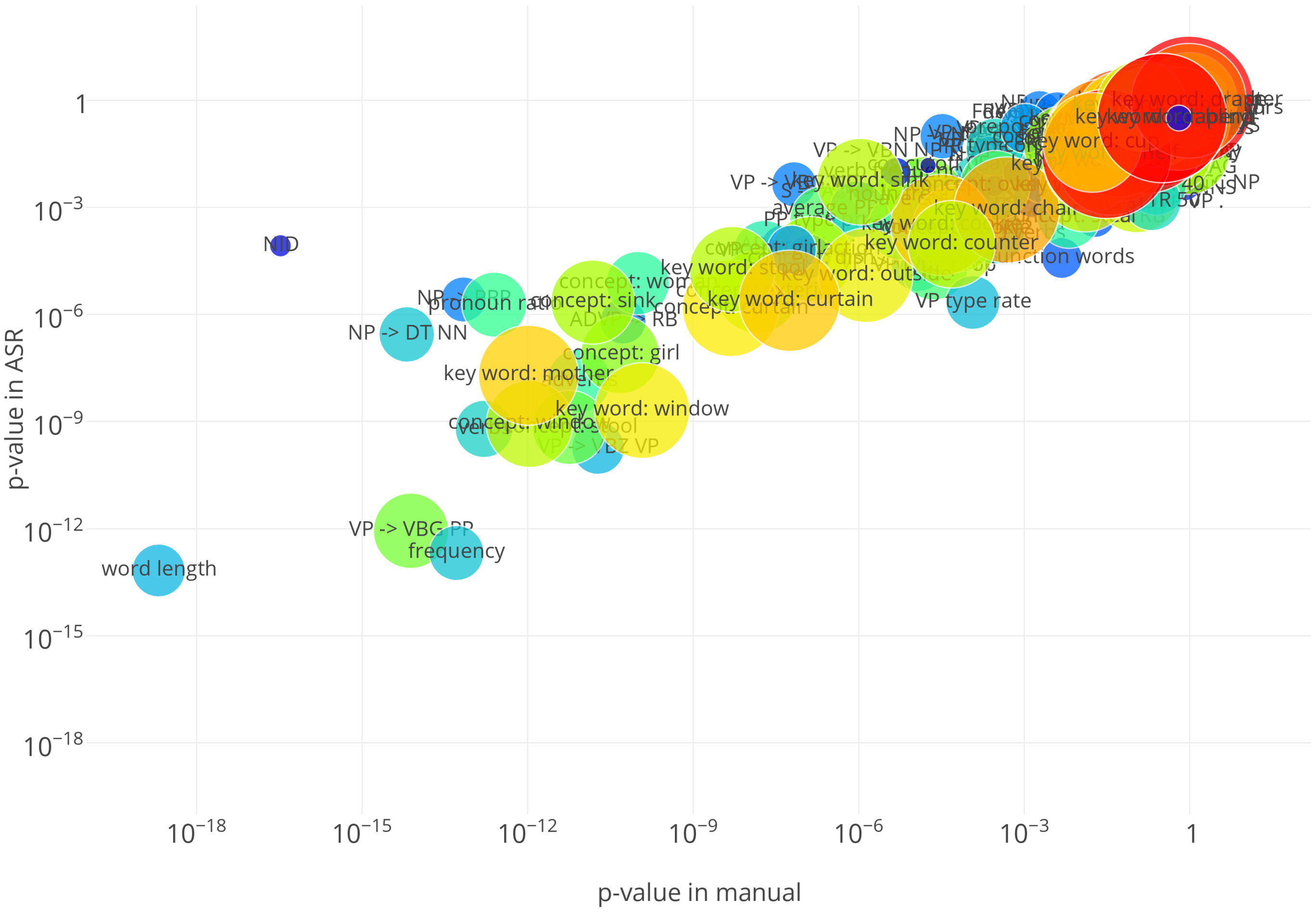


Figure 6.4: Bubbles represent features; large red bubbles indicate features that are highly correlated in the ASR and manual transcripts, while small blue bubbles indicate low correlation. Logarithmically scaled axes show features which are significant between the control and AD groups ($p \ll 1$).

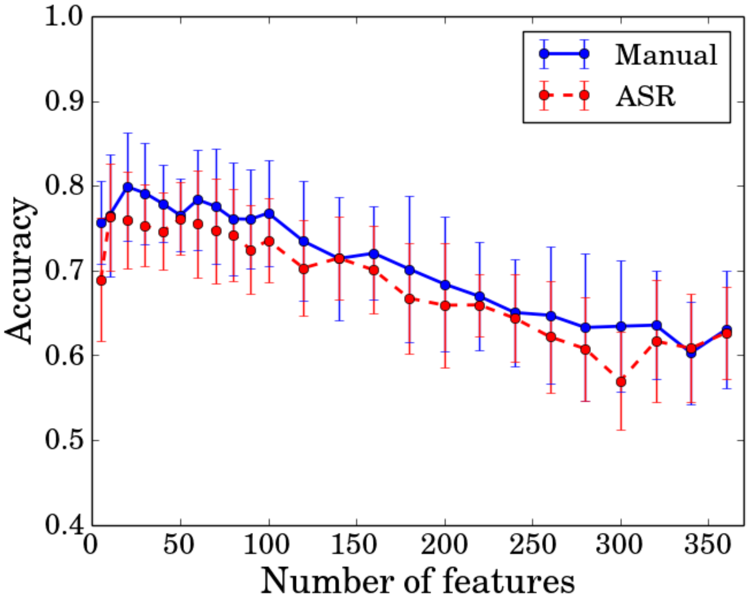
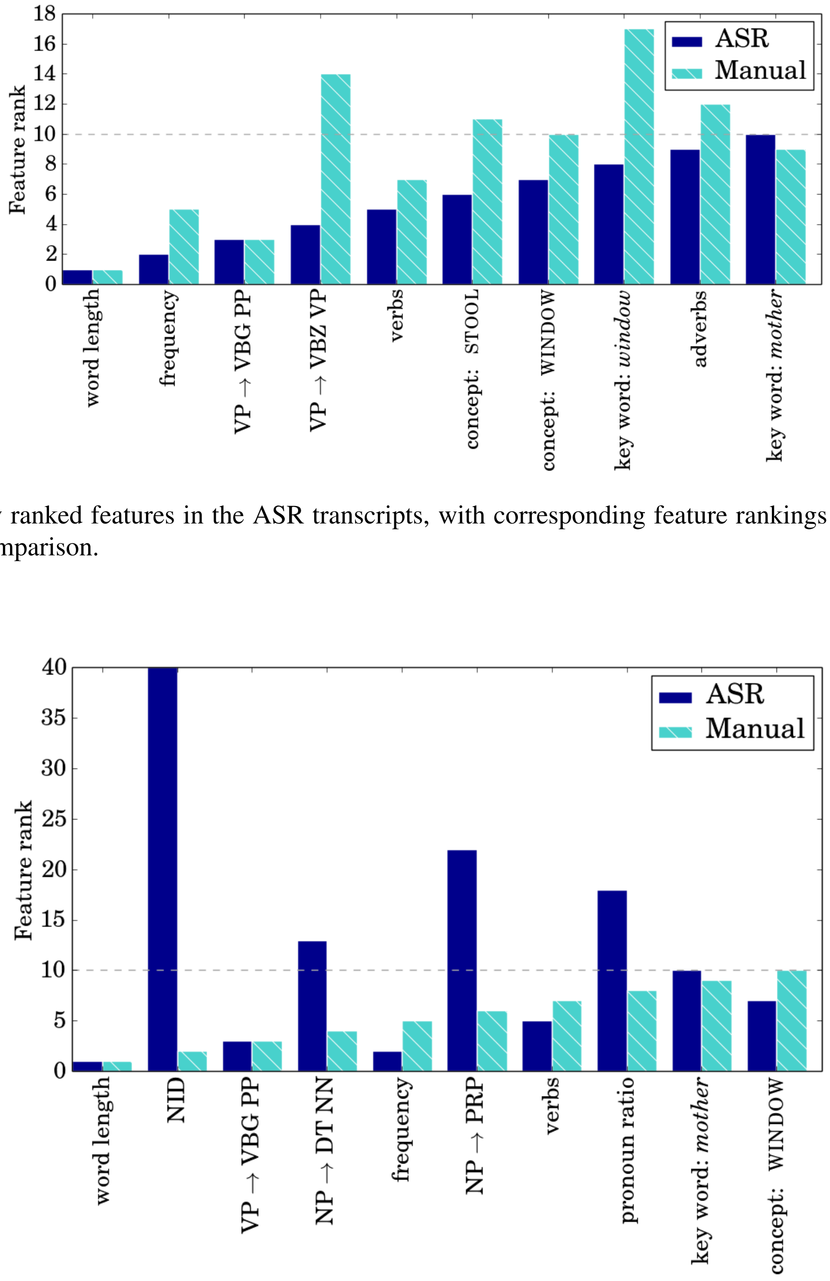


Figure 6.5: Average accuracies for distinguishing between AD and control narratives for varying feature sets. Error bars represent the standard deviation across folds.

6.4.2 Classification

We now compare the performance of a logistic regression classifier on the task of differentiating between AD and control narratives, using either the manual transcripts or the ASR transcripts. As in Section 6.3.2, we iterate over a number of different feature set sizes, selected using a correlation-based filter. The accuracy is averaged over 10 folds (restricted here to be the same folds used in the ASR process). The accuracies (and standard deviations) are shown in Figure 6.5. While the accuracies are lower on the ASR transcripts, as expected, they are within the error bars of the accuracies from the manual cases, except for when the number of features is very small, i.e., $N = 5$. The best accuracy for the ASR transcripts is 76.3%, achieved using 10 features. A very similar result is achieved using the manual transcripts with 10 features (76.5%).

To compare the features which were selected using the ASR and manual transcripts at $N = 10$, we rank all the features in each data set by their correlation with diagnosis, since this was the feature selection criterion. As before, the exact ranking will vary from fold to fold,



(a) Top 10 highly ranked features in the ASR transcripts, with corresponding feature rankings from the manual transcripts for comparison.

(b) Top 10 highly ranked features in the manual transcripts, with corresponding feature rankings from the ASR transcripts for comparison

Figure 6.6: Highly ranked features in the ASR and manual transcripts.

but this gives a global estimate of the relevance of each feature. Figure 6.6 shows the top 10 features in the ASR transcripts (with the corresponding rankings in the manual transcripts for comparison), and the top 10 features in the manual transcripts (with the corresponding rankings in the ASR transcripts).

In Figure 6.6a, we see that the top 10 features in the ASR transcripts are also highly ranked in the manual transcripts. Some of the features are ranked lower in the manual transcripts (e.g., VP → VBZ VP is ranked 14th rather than 4th). However, all of the top 10 features in the ASR transcripts are ranked in the top 17 in the manual transcripts, and we can see in Figure 6.4 that they are all statistically significant between the groups in both sets of transcripts.

In Figure 6.6b we see an anomalous result with the second-ranked feature, NID. As anticipated, NID is ranked much lower in the ASR transcripts than in the manual transcripts. Interestingly, NP → PRP (where PRP stands for 'personal pronoun') and pronoun ratio are both ranked lower in the ASR transcripts than in the manual transcripts. In Figure 6.4 these features are less significant (although still $p \sim 10^{-5}$) in the ASR transcripts than in the manual transcripts. Examining the values of these features, we find that NP → PRP has an average value of 0.059 for AD and 0.042 for controls in the manual transcripts, compared with 0.038 for AD and 0.028 for controls in the ASR transcripts. That is, the number of pronouns is underestimated in the ASR transcripts, but more so in the AD transcripts than the control transcripts, reducing the relative difference between the groups. A similar trend is observed for pronoun ratio. Cursory examination of the transcripts does not immediately suggest a reason for this effect, although future work will involve a more detailed analysis of these types of discrepancies.

In general, however, these plots suggest that the ASR transcripts still contain much of the same information that the manual transcripts do, with the exception of speech distortions that are coded as NID. Critically, the classifier is not making decisions based on features that are entirely irrelevant in the manual transcripts. Rather, there are relatively small differences in the order in which features are ranked, at least for the $N = 10$ case. Also worth noting is the fact that none of the acoustic features were selected in the top 10 in the ASR case, even though those features are not affected by the ASR errors.

6.4.3 Discussion

Using automatically generated transcripts, we were able to achieve classification accuracies that were within one standard deviation of the accuracies using manual transcriptions. While there is still room for improvement, this is a promising step towards the goal of a fully automated system. We also showed that there are a number of features which are robust to the ASR process and differ significantly between the participant groups. Unlike in our previous experiments with ASR (Section 4.1), these features were not limited to psycholinguistic features, but also included features measuring the occurrence of syntactic patterns and lexical keywords.

One limitation of this work is that we have no way of directly linking feature values between the ASR and manual transcripts. For example, it would be useful to be able to compare an ASR transcript to its original and see that the number of pronouns in the ASR transcript is reduced *because* some of them were deleted, and others were mistakenly substituted with determiners. This type of analysis will require aligning the two sets of transcripts and then associating features with insertion, substitution, and deletion errors. While not straightforward, this would potentially allow us to make improvements to the system as well as better interpret the measured feature values.

6.5 Summary

In this chapter, we have shown that the automated linguistic analysis of verbal picture descriptions can lead to high diagnostic classification accuracies in the case of Alzheimer's disease. This is particularly relevant given that language impairment is not part of the core criteria for AD, although changes to speech and language are widely reported in the literature. The success of the approach may be due to the fact that narrative speech production does not only involve the language centers of the brain, but also involves elements of planning, organization, and memory. Some researchers have suggested that other language elicitation tasks probe these cognitive skills more deeply; for example, asking the participant to describe everything they would need to do to prepare for a trip (Fleming, 2014). Further investigation will be required to see if analysis of other types of speech (e.g., conversational speech) lead to similar results.

Beyond the binary classification task, we also conducted a factor analysis, and identified four factors that explained most of the variance in the data: semantic impairment, syntactic impairment, information impairment, and acoustic abnormality. The factors were only loosely correlated, suggesting that individuals could, for example, show syntactic impairment without semantic impairment, and vice versa. Such dissociations have already been observed in PPA (and in aphasia in general), where more focal pathology is common. In the case of AD, we hypothesize that the heterogeneity in language symptoms is linked to the heterogenity of the underlying AD pathology, but are unable to say anything conclusively due to the absence of neuroimaging data for these patients. Associating features of spontaneous speech in AD with anatomical images would be a compelling project for future work, if such a data set should become available.

Finally, we repeated the classification experiment using automatically recognized transcripts and found that the accuracy was lower, but within one standard deviation, of the accuracy using manual transcripts. Furthermore, many of the relevant features in the manual transcripts were still highly significant in the ASR transcripts. This points to the feasibility of a fully automated processing pipeline for future applications.

Chapter 7

Conclusions and future work

7.1 Summary of contributions

In this dissertation, I have presented an in-depth computational analysis of connected speech in primary progressive aphasia (PPA), including the automated extraction of clinically-motivated measures of speech and language production, and the accurate classification of two PPA subtypes versus healthy controls, and versus each other. In the process, I have shown that traditional measures of syntactic complexity are not sensitive to the differences between the semantic variant of PPA (sv-PPA) and the nonfluent/agrammatic variant of PPA (nfv-PPA). As a result, I also extracted syntactic features measuring the frequency of different grammatical constituents, rather than simply the overall "complexity". These features were more discriminative between the two PPA subtypes, and detailed analysis revealed the somewhat fuzzy boundary between semantic and syntactic impairment. There were also numerous differences observed between sv-PPA and controls, suggesting that there can be changes to syntax in sv-PPA, in contrast to the traditional view and in agreement with other recent studies (Meteyard and Patterson, 2009; Meteyard et al., 2014). Including these additional features in an ablation study led to an increase in classification accuracy between the subtypes, from 79.2% to 91.7%.

I also augmented the feature set with dependency relation features, and was able to achieve a 76% classification accuracy between agrammatic PPA patients and agrammatic post-stroke aphasia patients, contributing to the ongoing debate regarding these two manifestations of agrammatism. Previous work has not agreed on whether the impairments seen in agrammatic PPA and agrammatic stroke are fundamentally different (Patterson et al., 2006), or the same (Thompson et al., 2013). Thompson et al. (2013) suggested that one explanation for the discrepancy could be different inclusion criteria; however, in the present work, participants were selected by the same criteria used by Thompson et al. (2013), eliminating that potential source of error. While the results here do suggest that differences between the groups exist, further work will be required to determine how well these findings generalize, and how they might relate to the underlying pattern of cortical atrophy in the two conditions.

Furthermore, I extended the approach to a different elicitation task (picture description rather than story-telling) and a different type of dementia (Alzheimer's disease, or AD). This involved introducing new features to measure repetitiveness and to assess the production of relevant information content. I was able to achieve 81% accuracy in differentiating between people with AD and healthy controls, as well as demonstrate that performance across the dimensions of semantics, syntax, information, and acoustics need not be correlated.

Finally, I attempted throughout this work to be mindful of the practical goal of a fully automated system. With the exception of Chapter 5, speech events such as restarts, filled pauses, comments about the task, asides, and other "non-narrative" speech were all included in the analysis, except where they could be removed automatically. I experimented with using automatically generated transcripts in both the PPA and AD cases, and presented a new way of visualizing the effect of ASR on the resulting features with the bubble plots in Section 6.4. I also considered the problem of boundary segmentation, which is a thorny issue in speech analysis even in the absence of ASR errors. I examined the effect of the automatically annotated boundaries on the traditional syntactic complexity metrics, which are particularly sensitive to the location of such boundaries. Somewhat surprisingly, I found that the classification accuracy using the automatically segmented transcripts was higher than using the manual transcripts, even though the features themselves were, in some cases, quite different in the two data sets. While indicating the feasibility of the automated approach, this also serves as a reminder that classification accuracy alone is not the whole story, and further supports the idea that critical analysis of the features can be valuable in interpreting the validity of the results.

Indeed, a minor contribution of this dissertation has been the ongoing exploration of how to compare feature values across groups, feature selection methods, and data sets. From the traditional approach of listing means and p values in a table (e.g., Table 3.4), to presenting lists of features with various font effects to indicate significance (Table 3.9), to more graphical methods such as selectivity ratios (Figure 3.2) and bar charts showing the distribution of feature values across groups (e.g., Figure 3.9), the proportion of folds in which a feature was selected (e.g., Figure 4.4), and the relative rankings of selected features (Figure 6.6), the process of visualizing these differences in a meaningful way has evolved over the duration of the research and is a subject of ongoing reflection and experimentation.

7.2 Limitations

One limitation of this work is the small size of the data sets, particularly for the PPA work. PPA is relatively rare, and clinical data in general tends to be scarce. In particular, to be given access to transcribed narrative speech data from PPA patients from two separate labs was an incredible stroke of luck (and generosity). Which is to say, the barrier to getting more data in this field is high, and thus improvements are not likely to come from training increasingly complex models on increasingly more data, as they have in some other applications of machine learning.

The main limitation of small data is that it is difficult to know how well these results generalize. Are we measuring properties of PPA speech in general, or just the idiosyncratic patterns of this particular group of people? I have tried to mitigate this concern in the following ways:

- By using significance testing, where appropriate, to identify statistically significant differences.

- By using a cross-validation training and testing framework for classification.

- By assessing how well the results fit into the literature and current understanding of PPA.

Using the extracted features to train and test a machine learning classifier is, in itself, a test of generalizability. If a feature is identified as important on the training set, but then shows no discriminative power in the test set, the subsequent classification accuracy will be low. However, this does not guard against the possibility that all or most of the participants in this small data set share some common linguistic quirks unrelated to their PPA diagnosis, or which appear simply by random chance.

This issue is somewhat less serious in the DementiaBank AD data set, which is an order of magnitude larger (though still small by machine learning standards). The bigger issue in the AD data is that the participant groups are not matched for age or education, factors which have been shown to affect performance on picture description tasks (e.g., Le Dorze and Bédard (1998)). While the differences in age (AD mean: 71.8, control mean: 65.2) and education (AD mean: 12.5, control mean: 14.2) are much smaller than the ranges considered in research such as that cited above, the issue reflects a real-world requirement to be able to factor age, gender, education, and other relevant demographic information into the analysis. One step towards meeting that requirement will be collecting normative data from people from these different demographic groups for comparison.

Another criticism of the AD data set is that the diagnoses were not made according to current diagnostic criteria. However, according to Becker et al. (1994), patients were diagnosed on the basis of their clinical history and their performance on neuropsychological testing, and the diagnoses were updated in 1992, taking into account any relevant information from the intervening years. Autopsies were performed on 50 patients, and in 43 cases the AD diagnosis was confirmed (86.0%). A more recent study of clinical diagnostic accuracy in AD found that of 526 cases diagnosed as probable AD, 438 were confirmed as neuropathological AD post-mortem (83.3%) Beach et al. (2012), suggesting that the DementiaBank diagnoses are generally as reliable as diagnoses made using present-day criteria.

One finding that repeatedly emerged from this work was the importance of the psycholinguistic measures of frequency, familiarity, imageability, and age-of-acquisition. Our frequency ratings were calculated from a large database of film and television subtitles (Brysbaert and New, 2009) and covered most of the words in our PPA corpus. However, the other three metrics are subjective, and the norms for these quantities were compiled via undergraduate questionnaires (Stadthagen-Gonzalez and Davis, 2006; Gilhooly and Logie, 1980). These norms contained a combined 3,394 words, and covered only around 30% of the words in our PPA corpus. In future work, we could benefit from new approaches to estimating these measures. One potential path lies in crowd-sourcing. For example, Kuperman et al. (2012) recently presented a crowd-sourced database of age-of-acquisition ratings for 30,000 words. In a different approach, recent work has used existing norms as training data for supervised learning methods to automatically predict ratings for familiarity, age-of-acquisition, concreteness and imageability, achieving a correlation of .88 with human judgement (Paetzold and Specia, 2016). Related work in German generated estimates for the abstractness, arousal, imageability and emotional valence associated with 350,000 words (Köper and Schulte im Walde, 2016). By using similar techniques, we could increase the coverage of the norms on our data and improve our estimates for these psycholinguistic measures.

Given that analysis of the selected features has been so integral to the work, another aspect which could potentially be improved upon is the feature selection. In the methodology considered here, the feature selection step is entirely separate from the classification step, and the feature selection method is typically a ‘filter’ method based on p values or correlation coefficients. The benefit of this architecture is that feature selection and classification algorithms can easily be swapped in and out. However, Guyon and Elisseeff (2003) give examples of how information can be lost using these simple filter methods. For example, two features which are useless on their own can provide predictive information when they are taken together. These relationships are not captured when features are assessed individually. In contrast, 'wrapper' methods optimize classifier performance directly by selecting exactly those features which lead to the best performance by the classifier on the training set. The drawback to wrapper methods is that the process of identifying the optimal set of features can be computationally intractable. However, such methods could potentially lead to better accuracies, reduce redundancy in the feature sets, and capture information from complex relationships between the features.

A related limitation involves the interpretability of the machine learning models themselves. We have essentially treated the classifier in each experiment as a 'black box' which takes features in and outputs class labels. Rüping (2006) argues that feature selection in combination with a black box classifier represents only the first level of interpretability in machine learning. Interpretability can be increased by examining how the classifier generates the output from the input. A classic example of this is the decision tree classifier, whose final model is relatively easy for the user to understand, and may even model the human decision-making process in some applications. Rüping also suggests that when the number of features is large, interpretability can be improved by training small, local models to improve on the performance of more easily understood (but perhaps less accurate) global models. The implementation of similar methods here could further increase the clinical utility of the work.

7.3 Future work

Since dementia is progressive, in many ways it makes sense to predict severity rather than binary class membership. A first step towards this goal could involve modifying the existing pipeline to perform regression, rather than classification; for example, using logistic regression to predict MMSE scores rather than diagnosis in the DementiaBank corpus. This would allow us to assign a severity rating to any given narrative. However, being able to predict severity would also allow longitudinal monitoring of the sort discussed in Section 2.7.3. Of particular interest would be developing a method to generate aggregate scores for different linguistic areas (in the same spirit as the factor analysis of Section 6.3.4), which could then be monitored over time. Monitoring patterns of language decline in dementia could help track the spread of cortical atrophy. However, such technology could also be applied to other language disorders with a more positive prognosis, such as post-stroke aphasia and childhood language disorders. In such cases, longitudinal severity ratings across specific language areas could provide a quantitative basis for evaluating the efficacy of potential therapies, as well as suggesting particular language deficits that should be intensively targeted.

One focus of this dissertation has been the development of metrics which are sensitive to *syntactic* changes in language. In future work, I would like to investigate more sophisticated measures of *semantic* production. In particular, the keyword-spotting approach to assessing information content and lexical choice in Chapter 6 was relatively naïve, and in the PPA work I did not directly consider the semantic content of the narratives at all. In recent years, there has been growing research into vector space models of semantics, and the methods being developed in that area could prove to be valuable here. If we construct a semantic space using a corpus of Alzheimer's speech, how will that differ from a representation trained on data from healthy speakers? Furthermore, can this tell us anything at all about how semantic knowledge is represented in the brain? Others have proposed the idea of "conditional word similarity" based on the premise that different populations will tend to use a given word in different contexts (Kiros et al., 2014). We may be able to use such concepts to detect subtle differences in semantic processing, either compared to a control model or, ideally, to a model built on an individual's language use earlier in life.

Beyond syntax and semantics, higher level discourse processing is also an area of interest. For example, Ash et al. (2006) found that some dementia patients who were not frankly aphasic still had difficulty organizing their narratives and making connections between story themes. Previous work in computational linguistics has suggested automated methods for computing discourse measures such as cohesion (Graesser et al., 2004) and coherence (Lapata and Barzilay, 2005), which may help to reveal deficits in planning and organization not captured by our current system.

In the work presented here, speech data were elicited through specific story-telling and picture-description tasks. The relatively constrained nature of these tasks (compared to open-ended conversation) makes it easier to evaluate the content of the narratives, and compare performance across participants. However, the tasks are somewhat artificial, and previous work has suggested that an individual's performance on such tasks can very from day to day (Cook et al., 2009). Other research has suggested that conversational speech may offer a more realistic perspective of language performance (Coelho et al., 2003). Conversational analysis would also provide an opportunity to assess language use in a social context, including behaviours such as turn-taking, back-channelling, lexical entrainment, and potentially even non-linguistic elements of communication like gesture and eye contact.

Many of the available pharmacological interventions for Alzheimer's disease symptoms are most effective if administered early in the disease progression (Solomon and Murphy, 2005). Thus, identifying and accurately describing the early linguistic symptoms of AD is critical. Semi-structured speech data of the type discussed here is not normally collected until after an individual starts displaying obvious symptoms and visits a clinician. Furthermore, as discussed previously, the average person does not have speech samples (or novels, for that matter) which may be retrospectively analyzed to search for relevant linguistic changes. However, in recent years it has become more common for people to generate large quantities of digital text, much of which is stored indefinitely on the Internet. Therefore, a long-term research goal will be to develop methods of mining alternative sources of language data, including digital text such as email, text messages, social media posts, and blogs. While perhaps not quite as spontaneous as speech, such written material is often generated quickly and without the help of external sources, such as dictionaries or copy editors. This research will study questions about how language disorders are expressed in digital media, and how that expression differs from the usual non-standard grammar and spelling frequently seen among cognitively healthy users. It will also allow us to explore what compensatory strategies are used by people with incipient language or cognitive disorders in their digital communication (predictive spelling, for example), and how they could be implemented in assistive technologies.

I am also interested in combining language analysis with other information about a person's cognitive status. One example of additional information is neuropsychological test scores. This could include global measures of cognitive decline, such as the MMSE (as mentioned above), but another possibility would be to examine the correlations between specific language tests and narrative speech measures. Do formal language testing and narrative speech analysis measure the same things (in which case, one method could be considered redundant), or is there information that can be obtained only from one method or the other? If the latter, can we combine these sources of information to improve diagnostic accuracy? Another potential source of information is neuroimaging. Correlations between different brain regions and the ability to produce and comprehend language have long been known, but there are still many open questions about the interactions between these areas and other parts of the brain. Being able to associate specific patterns of language decline with patterns of degeneration could contribute to our understanding of language processing in the brain (see Wilson et al. (2011) for one example of this in PPA), as well as provide a non-invasive biomarker for neurodegeneration. As a third possibility, analysis of video data in combination with speech data could also hold promise for future work. Previous work has identified changes in the eye movements (Anderson and MacAskill, 2013), facial expressions (Seidl et al., 2012), and gestures (Carlomagno et al., 2005) of people with dementia. A benefit to combining information from multiple input modalities may be increased sensitivity to early signs of dementia; for example, Alberdi et al. (2016) argue that a multimodal approach will be crucial for early diagnosis, in order to detect small physiological and behavioural changes.

7.4 Concluding remarks

From a computational linguistics perspective, I have demonstrated how we can transfer techniques from solutions to problems as disparate as authorship attribution, language learning, and grammar checking to tackle problems in the healthcare space. I have argued that the features we extract need not be viewed simply as a means for increasing classification accuracy. Rather, they can tell us useful things about the data, and in this particular case, about how people with dementia use language. From the perspective of dementia and aphasia research, I have shown that computer analysis can reveal much of the same information that has previously been extracted by hand, as well as providing new ways of exploring the data that would not be reasonable to attempt manually. While acknowledging the limitations of the current work and the challenges that lie ahead, I have presented multiple experiments showing that we can detect a useful signal in noisy data, and suggested that fully automated analyses may have some practical applications in the future. While this dissertation has centered on dementia, it is easy to imagine how similar ideas could be applied in related healthcare fields where changes to speech and language must be detected and assessed: traumatic brain injury, childhood language disorders, depression, schizophrenia, etc. Such work will necessarily be multi-disciplinary, and will be enriched by expertise from domain experts in fields such as medicine, psychology, and speech-language pathology. My hope is that we in computational linguistics can ultimately further knowledge in those fields by developing new techniques to search for and understand patterns in the noisy reality of natural language.

> "The job of the linguist, like that of the biologist or the botanist, is not to tell us how nature should behave, or what its creations should look like, but to describe those creations in all their messy glory and try to figure out what they can teach us about life, the world, and, especially in the case of linguistics, the workings of the human mind."
>
> –Arika Okrent

Bibliography

Abney, S. (1990). Rapid incremental parsing with repair. In *Proceedings of the 6th New OED Conference: Electronic Text Research*, pages 1–9.

Adlam, A.-L. R., Bozeat, S., Arnold, R., Watson, P., and Hodges, J. R. (2006). Semantic knowledge in mild cognitive impairment and mild Alzheimer's disease. *Cortex*, 42(5):675–684.

Adnène, C., Lamia, B., and Mounir, M. (2003). Analysis of pathological voices by speech processing. In *Proceedings of the 7th International Symposium on Signal Processing and its Applications*, pages 365–367. IEEE.

Agresti, A. (2003). *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Ahmed, S., de Jager, C. A., Haigh, A.-M. F., and Garrard, P. (2012). Logopenic aphasia in Alzheimer's disease: Clinical variant or clinical feature? *Journal of Neurology, Neurosurgery & Psychiatry*, pages 1056–1062.

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737.

Alberdi, A., Aztiria, A., and Basarab, A. (2016). On the early diagnosis of Alzheimer's disease from multimodal signals: A survey. *Artificial Intelligence in Medicine*. In Press.

Alzheimer's Association (2016). 2016 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4):459–509.

Aman, F., Vacher, M., Rossato, S., and Portet, F. (2013). Analysing the performance of automatic speech recognition for ageing voice: Does it correlate with dependency level? In *Proceedings of the 4th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 9–15.

Amici, S., Gorno-Tempini, M. L., Ogar, J. M., Dronkers, N. F., and Miller, B. L. (2006). An overview on primary progressive aphasia and its variants. *Behavioural Neurology*, 17(2):77–87.

Anderson, T. J. and MacAskill, M. R. (2013). Eye movements in patients with neurodegenerative disorders. *Nature Reviews Neurology*, 9(2):74–85.

Appell, J., Kertesz, A., and Fisman, M. (1982). A study of language functioning in Alzheimer patients. *Brain and Language*, 17(1):73–91.

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Ash, S., McMillan, C., Gunawardena, D., Avants, B., Morgan, B., Khan, A., Moore, P., Gee, J., and Grossman, M. (2010). Speech errors in progressive non-fluent aphasia. *Brain and Language*, 113(1):13–20.

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

Ash, S., Moore, P., Vesely, L., Gunawardena, D., McMillan, C., Anderson, C., Avants, B., and Grossman, M. (2009). Non-fluent speech in frontotemporal lobar degeneration. *Journal of Neurolinguistics*, 22(4):370–383.

Ashford, J. W., Borson, S., O'Hara, R., Dash, P., Frank, L., Robert, P., Shankle, W. R., Tierney, M. C., Brodaty, H., Schmitt, F. A., Kraemer, H. C., and Buschke, H. (2006). Should older adults be screened for dementia? *Alzheimer's & Dementia*, 2(2):76–85.

Ashford, J. W., Borson, S., O'Hara, R., Dash, P., Frank, L., Robert, P., Shankle, W. R., Tierney, M. C., Brodaty, H., Schmitt, F. A., Kraemer, H. C., Buschke, H., and Fillit, H. (2007). Should older adults be screened for dementia? It is important to screen for evidence of dementia! *Alzheimer's & Dementia*, 3(2):75–80.

Baayen, H., Van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Baghai-Ravary, L. and Beet, S. W. (2012). *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*. Springer Science & Business Media.

Barney, A., Nikolic, D., Nemes, V., and Garrard, P. (2013). Detecting repeated speech: a possible marker for Alzheimer's disease. In *Proceedings of the 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 31–32.

Bastiaanse, R. and Thompson, C. K. (2003). Verb and auxiliary movement in agrammatic Broca's aphasia. *Brain and Language*, 84(2):286–305.

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., and Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *The Clinical Neuropsychologist*, 26(2):177–196.

Bayles, K. A., Tomoeda, C. K., McKnight, P. E., Helm-Estabrooks, N., and Hawley, J. N. (2004). Verbal perseveration in individuals with Alzheimer's disease. *Seminars in Speech and Language*, 26(04):335–347.

Beach, T. G., Monsell, S. E., Phillips, L. E., and Kukull, W. (2012). Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *Journal of Neuropathology & Experimental Neurology*, 71(4):266–273.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Benedet, M., Patterson, K., Gomez-Pastor, I., and Garcia de la Rocha, M. L. (2006). 'Non-semantic' aspects of language in semantic dementia: As normal as they're said to be? *Neurocase*, 12(1):15–26.

Berisha, V., Wang, S., LaCross, A., and 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*, 45(3):959–963.

Berndt, R. S., Haendiges, A. N., Mitchum, C. C., and Sandson, J. (1997). Verb retrieval in aphasia: Relationship to sentence processing. *Brain and Language*, 56(1):107–137.

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E., and Schwartz, M. (2000). *Quantitative Production Analysis: A Training Manual for the Analysis of Aphasic Sentence Production*. Psychology Press, Hove, UK.

Bethard, S. J. (2007). *Finding event, temporal and causal structure in text: A machine learning approach*. PhD thesis, University of Colorado.

Billette, O. V., Sajjadi, S. A., Patterson, K., and Nestor, P. J. (2015). SECT and MAST: New tests to assess grammatical abilities in primary progressive aphasia. *Aphasiology*, 29(10):1135–1151.

Bird, H. and Franklin, S. (1996). Cinderella revisited: A comparison of fluent and non-fluent aphasic speech. *Journal of Neurolinguistics*, 9(3):187–206.

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

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

Bishop, D. V. M. (2003). *Test for the Reception of Grammar (TROG-2) Version 2*. Psychological Corporation, London.

Blonder, L. X., Kort, E. D., and Schmitt, F. A. (1994). Conversational discourse in patients with Alzheimer's disease. *Journal of Linguistic Anthropology*, 4(1):50–71.

Boustani, M., Peterson, B., Hanson, L., Harris, R., and Lohr, K. N. (2003). Screening for dementia in primary care: A summary of the evidence for the US Preventive Services Task Force. *Annals of Internal Medicine*, 138(11):927–937.

Breedin, S. D., Saffran, E. M., and Schwartz, M. F. (1998). Semantic factors in verb retrieval: An effect of complexity. *Brain and Language*, 63:1–31.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bridges, K. A. and Van Lancker Sidtis, D. (2013). Formulaic language in Alzheimer's disease. *Aphasiology*, pages 1–12.

Brooke, J. and Hirst, G. (2012). Robust, lexicalized native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 391–408.

Brookes, M. (1997). Voicebox: Speech processing toolbox for Matlab. `www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

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

Calzà, L., Beltrami, D., Gagliardi, G., Ghidoni, E., Marcello, N., Rossini-Favretti, R., and Tamburini, F. (2015). Should we screen for cognitive decline and dementia? *Maturitas*, 82(1):28–35.

Carlomagno, S., Pandolfi, M., Marini, A., Di Iasi, G., and Cristilli, C. (2005). Coverbal gestures in Alzheimer's type dementia. *Cortex*, 41(4):535–546.

Chae, J. and Nenkova, A. (2009). Predicting the fluency of text with shallow structural features: Case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 139–147.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 132–139.

Chen, M. and Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 722–731.

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

Chomsky, N. (1957). *Syntactic Structures*. Mouton de Gruyter, Berlin.

Clark, H. H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Coelho, C., Youse, K., Le, K., and Feinn, R. (2003). Narrative and conversational discourse of adults with closed head injuries and non-brain-injured adults: A discriminant analysis. *Aphasiology*, 17(5):499–510.

Cook, C., Fay, S., and Rockwood, K. (2009). Symptom fluctuation in patients with Alzheimer's disease in the VISTA clinical trial. *Canadian Journal of Geriatrics*, 12:177–182.

Costello, A. and Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7):1–9.

Cotelli, M., Borroni, B., Manenti, R., Alberici, A., Calabria, M., Agosti, C., Arevalo, A., Ginex, V., Ortelli, P., Binetti, G., Zanetti, O., Padovani, A., and Cappa, S. F. (2006). Action and object naming in frontotemporal dementia, progressive supranuclear palsy, and corticobasal degeneration. *Neuropsychology*, 20(5):558–565.

Covington, M. A. and McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., and Trillet, M. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and language*, 53(1):1–19.

Croot, K., Patterson, K., and Hodges, J. R. (1998). Single word production in nonfluent progressive aphasia. *Brain and Language*, 61(2):226–273.

Cuendet, S., Shriberg, E., Favre, B., Fung, J., and Hakkani-Tür, D. (2007). An analysis of sentence segmentation features for broadcast news, broadcast conversations, and meetings. In *Proceedings of the Workshop on Searching Spontaneous Conversational Speech*, pages 43–49.

Cuetos, F., Arango-Lasprilla, J. C., Uribe, C., Valencia, C., and Lopera, F. (2007). Linguistic changes in verbal expression: A preclinical marker of Alzheimer's disease. *Journal of the International Neuropsychological Society*, 13(3):433–439.

Cummings, J. L., Benson, D. F., Hill, M. A., and Read, S. (1985). Aphasia in dementia of the Alzheimer type. *Neurology*, 35(3):394–394.

Cummins, N., Epps, J., Breakspear, M., and Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. In *Proceedings of INTERSPEECH*, pages 2997–3000.

da Cunha, A. L. V., de Sousa, L. B., Mansur, L. L., and Aluisio, S. M. (2015). Automatic proposition extraction from dependency trees: Helping early prediction of Alzheimer's disease from narratives. In *Proceedings of the IEEE 28th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 127–130.

D'Arcy, S., Rapcan, V., Penard, N., Morris, M. E., Robertson, I. H., and Reilly, R. B. (2008). Speech as a means of monitoring cognitive function of elderly speakers. In *Proceedings of INTERSPEECH*, pages 2230–2233.

de Lira, J. O., Ortiz, K. Z., Campanha, A. C., Bertolucci, P. H. F., and Minett, T. S. C. (2011). Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics*, 23(03):404–412.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, volume 6, pages 449–454.

Dede, E., Zalonis, I., Gatzonis, S., and Sakas, D. (2015). Integration of computers in cognitive assessment and level of comprehensiveness of frequently used computerized batteries. *Neurology, Psychiatry and Brain Research*, 21(3):128–135.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1):4–34.

Druks, J. and Carroll, E. (2005). The crucial role of tense for verb production. *Brain and Language*, 94(1):1–18.

Dunn, L. M. and Dunn, L. M. (1997). *Peabody Picture Vocabulary Test*. American Guidance Service, Circle Pines, Minnesota, 3rd edition.

Duong, A., Giroux, F., Tardif, A., and Ska, B. (2005). The heterogeneity of picture-supported narratives in Alzheimer's disease. *Brain and Language*, 93(2):173–184.

Ehrlich, J. S., Obler, L. K., and Clark, L. (1997). Ideational and semantic contributions to narrative production in adults with dementia of the Alzheimer's type. *Journal of Communication Disorders*, 30(2):79–99.

Elvevåg, B., Foltz, P. W., Rosenstein, M., and DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics*, 23(3):270–284.

Emmorey, K. D. (1987). The neurological substrates for prosodic aspects of speech. *Brain and Language*, 30(2):305–320.

Erard, M. (2008). *Um...: slips, stumbles, and verbal blunders, and what they mean*. Anchor Canada.

Evang, K., Basile, V., Chrupala, G., and Bos, J. (2013). Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1422–1426.

Faber-Langendoen, K., Morris, J. C., Knesevich, J. W., LaBarge, E., Miller, J. P., and Berg, L. (1988). Aphasia in senile dementia of the Alzheimer type. *Annals of Neurology*, 23(4):365–370.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272–299.

Faroqi-Shah, Y. and Thompson, C. K. (2007). Verb inflections in agrammatic aphasia: Encoding of tense features. *Journal of Memory and Language*, 56(1):129–151.

Feldman, H. and Woodward, M. (2005). The staging and assessment of moderate to severe Alzheimer disease. *Neurology*, 65(6):S10–S17.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Fergadiotis, G. and Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology*, 25(11):1414–1430.

Ferris, S. H. and Farlow, M. (2013). Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. *Clinical Interventions in Aging*, 8:1007–1014.

Fleming, V. B. (2014). Early detection of cognitive-linguistic change associated with mild cognitive impairment. *Communication Disorders Quarterly*, 35(3):146–157.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12:189–198.

Forbes, K. E., Venneri, A., and Shanks, M. F. (2001). Distinct patterns of spontaneous speech deterioration: An early predictor of Alzheimer's disease. *Brain and Cognition*, 48(2):356–361.

Forbes-McKay, K., Shanks, M. F., and Venneri, A. (2013). Profiling spontaneous speech decline in Alzheimer's disease: A longitudinal study. *Acta Neuropsychiatrica*, 25(06):320–327.

Forbes-McKay, K. E. and Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences*, 26:243–254.

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3):354–375.

Francois, K. (2008). The comprehensive Dragon NaturallySpeaking guide. Presentation at the Inclusive Learning Technologies Conference.

Fraser, K., Rudzicz, F., Graham, N., and Rochon, E. (2013a). Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pages 47–54.

Fraser, K. C., Ben-David, N., Hirst, G., Graham, N. L., and Rochon, E. (2015a). Sentence segmentation of aphasic speech. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 862–871.

Fraser, K. C., Hirst, G., Graham, N. L., Meltzer, J. A., Black, S. E., and Rochon, E. (2014a). Comparison of different feature sets for identification of variants in progressive aphasia. In *Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 17–26.

Fraser, K. C., Hirst, G., Meltzer, J. A., Mack, J. E., and Thompson, C. K. (2014b). Using statistical parsing to detect agrammatic aphasia. In *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 134–142.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., and Rochon, E. (2014c). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2015b). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Fraser, K. C., Rudzicz, F., and Rochon, E. (2013b). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Proceedings of INTERSPEECH*, pages 2177–2181.

Funnel, E. (1996). W.L.P.: A case for the modularity of language function and dementia. In Code, C., Wallesch, C.-W., Joanette, Y., and Lecours, A. R., editors, *Classic Cases in Neuropsychology*. Psychology Press, Hove, UK.

Gabani, K., Solorio, T., Liu, Y., Hassanali, K.-n., and Dollaghan, C. A. (2011). Exploring a corpus-based approach for detecting language impairment in monolingual English-speaking children. *Artificial Intelligence in Medicine*, 53(3):161–170.

Garrard, P. and Forsyth, R. (2010). Abnormal discourse in semantic dementia: A data-driven approach. *Neurocase*, 16(6):520–528.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., and Gorno-Tempini, M. L. (2014). Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*, 55:122–129.

Gates, N. J. and Kochan, N. A. (2015). Computerized and on-line neuropsychological testing for late-life cognition and neurocognitive disorders: Are we there yet? *Current Opinion in Psychiatry*, 28(2):165–172.

Gavalda, M., Zechner, K., and Aist, G. (1997). High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 12–15.

Giles, E., Patterson, K., and Hodges, J. R. (1996). Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: Missing information. *Aphasiology*, 10(4):395–408.

Gilhooly, K. and Logie, R. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods*, 12:395–427.

Glosser, G. and Deser, T. (1991). Patterns of discourse production among neurological patients with fluent language disorders. *Brain and language*, 40(1):67–88.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Goodglass, H. (1993). *Understanding Aphasia*. Academic Press, Inc., San Diego.

Goodglass, H., Christiansen, J. A., and Gallagher, R. E. (1994). Syntactic constructions used by agrammatic speakers: Comparison with conduction aphasics and normals. *Neuropsychology*, 8(4):598.

Goodglass, H. and Kaplan, E. (1983). *The Assessment of Aphasia and Related Disorders*. Lea and Febiger, Philadelphia, Pennsylvania, 2nd edition.

Gorno-Tempini, M. L., Dronkers, N. F., Rankin, K. P., Ogar, J. M., Phengrasamy, L., Rosen, H. J., Johnson, J. K., Weiner, M. W., and Miller, B. L. (2004). Cognition and anatomy in three variants of primary progressive aphasia. *Annals of Neurology*, 55(3):335–346.

Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., Ogar, J. M., Rohrer, J. D., Black, S., Boeve, B. F., Manes, F., Dronkers, N. F., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B. L., Knopman, D. S., Hodges, J. R., Mesulam, M. M., and Grossman, M. (2011). Classification of primary progressive aphasia and its variants. *Neurology*, 76:1006–1014.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Graff, D. and Cieri, C. (2003). *English Gigaword Corpus*. Linguistic Data Consortium.

Graham, N. L., Patterson, K., and Hodges, J. R. (2004). When more yields less: Speaking and writing deficits in nonfluent progressive aphasia. *Neurocase*, 10(2):141–155.

Grossman, M. (2010). Primary progressive aphasia: Clinicopathological correlations. *Nature Reviews Neurology*, 6:88–97.

Guinn, C. I. and Habash, A. (2012). Language analysis of speakers with dementia of the Alzheimer's type. In *AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*, pages 8–13.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Haenlein, M. and Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics*, 3(4):283–297.

Hakkani-Tür, D., Vergyri, D., and Tür, G. (2010). Speech-based automated cognitive status assessment. In *Proceedings of INTERSPEECH*, pages 258–261.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutmann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations*, 2(1).

Harciarek, M. and Kertesz, A. (2011). Primary progressive aphasias and their contribution to the contemporary knowledge about the brain-language relationship. *Neuropsychology Review*, 21:271–287.

Hassanali, K.-N., Liu, Y., Iglesias, A., Solorio, T., and Dollaghan, C. (2013). Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods*, pages 1–9.

Heitkamp, N., Schumacher, R., Croot, K., de Langen, E. G., Monsch, A. U., Baumann, T., and Danek, A. (2016). A longitudinal linguistic analysis of written text production in a case of semantic variant primary progressive aphasia. *Journal of Neurolinguistics*, 39:26–37.

Henry, J. D., Crawford, J. R., and Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia*, 42(9):1212–1222.

Herrmann, N., Lanctôt, K. L., and Hogan, D. B. (2013). Pharmacological recommendations for the symptomatic treatment of dementia: The Canadian Consensus Conference on the Diagnosis and Treatment of Dementia 2012. *Alzheimer's Research and Therapy*, 5(Suppl. 1):1–12.

Hillis, A. E., Oh, S., and Ken, L. (2004). Deterioration of naming nouns versus verbs in primary progressive aphasia. *Annals of Neurology*, 55(2):268–275.

Hodges, J. R. and Patterson, K. (1995). Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. *Neuropsychologia*, 33(4):441–459.

Hunt, K. W. (1966). Recent measures in syntactic development. *Elementary English*, 43(7):732–739.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pages 27–36.

Jarrold, W., Peintner, B., Yeh, E., Krasnow, R., Javitz, H., and Swan, G. (2010). Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer's disease. In Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., and Huang, J., editors, *Brain Informatics*, volume 6334 of *Lecture Notes in Computer Science*, pages 299–307. Springer Berlin / Heidelberg.

Jørgensen, F. (2007). Clause boundary detection in transcribed spoken language. In *Proceedings from the Nordic Conference of Computational Linguistics (NODALIDA)*, pages 235–239.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2nd edition.

Jurica, P. J., Leitten, C. L., and Mattis, S. (2001). *Dementia Rating Scale-2*. Psychological Assessment Resources, Inc., Lutz, Florida.

Kaplan, E., Goodglass, H., and Weintraub, S. (2001). *Boston Naming Test*. Lippincott Williams & Wilkins, Philadelphia, 2nd edition.

Karam, Z. N., Provost, E. M., Singh, S., Montgomery, J., Archer, C., Harrington, G., and Mcinnis, M. G. (2014). Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4858–4862.

Kavé, G., Leonard, C., Cupit, J., and Rochon, E. (2007). Structurally well-formed narrative production in the face of severe conceptual deterioration: A longitudinal case study of a woman with semantic dementia. *Journal of Neurolinguistics*, 20(2):161–177.

Kaye, J. (2008). Home-based technologies: A new paradigm for conducting dementia prevention trials. *Alzheimer's & Dementia*, 4(1):S60–S66.

Kemper, S., Thompson, M., and Marquis, J. (2001). Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging*, 16(4):600–614.

Kempler, D. (1995). Language changes in dementia of the Alzheimer type. *Dementia and Communication*, pages 98–114.

Kempler, D., Curtiss, S., and Jackson, C. (1987). Syntactic preservation in Alzheimer's disease. *Journal of Speech, Language, and Hearing Research*, 30(3):343–350.

Kent, R. D. and Kim, Y.-J. (2003). Toward an acoustic typology of motor speech disorders. *Clinical linguistics & phonetics*, 17(6):427–445.

Kertesz, A. (1982). *Western Aphasia Battery Test Manual*. Psychological Corporation.

Kim, M. and Thompson, C. K. (2000). Patterns of comprehension and production of nouns and verbs in agrammatism: Implications for lexical organization. *Brain and Language*, 74(1):1–25.

Kiros, R., Zemel, R., and Salakhutdinov, R. R. (2014). A multiplicative model for learning distributed text-based attribute representations. *Advances in Neural Information Processing Systems*, pages 2348–2356.

Kirshner, H. S. (2012). Primary progressive aphasia and Alzheimer's disease: Brief history, recent evidence. *Current Neurology and Neuroscience Reports*, 12(6):709–714.

Kirshner, H. S., Webb, W. G., and Kelly, M. P. (1984). The naming disorder of dementia. *Neuropsychologia*, 22(1):23–30.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL)*, pages 423–430.

Knibb, J. A., Woollams, A. M., Hodges, J. R., and Patterson, K. (2009). Making sense of progressive non-fluent aphasia: An analysis of conversational speech. *Brain*, 132(10):2734–2746.

Kolář, J. (2008). *Automatic segmentation of speech into sentence-like units*. PhD thesis, University of West Bohemia.

Kolár, J., Liu, Y., and Shriberg, E. (2009). Genre effects on automatic sentence segmentation of speech: A comparison of broadcast news and broadcast conversations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4701–4704.

Köper, M. and Schulte im Walde, S. (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Kvalheim, O. M. (2010). Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *Journal of Chemometrics*, 24(7–8):496–504.

Labský, M., Cuřín, J., Macek, T., Kleindienst, J., Kunc, L., Young, H., Thyme-Gobbel, A., and Quast, H. (2012). Impact of word error rate on driving performance while dictating short texts. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 179–182. ACM.

Lambon Ralph, M. A., Graham, K. S., Ellis, A. W., and Hodges, J. R. (1998). Naming in semantic dementia—what matters? *Neuropsychologia*, 36(8):775–784.

Lambon Ralph, M. A., Patterson, K., Garrard, P., and Hodges, J. R. (2003a). Semantic dementia with category specificity: A comparative case-series study. *Cognitive Neuropsychology*, 20(3–6):307–326.

Lambon Ralph, M. A., Patterson, K., Graham, N., Dawson, K., and Hodges, J. R. (2003b). Homogeneity and heterogeneity in mild cognitive impairment and Alzheimer's disease: A cross-sectional and longitudinal study of 55 cases. *Brain*, 126(11):2350–2362.

Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 5, pages 1085–1090.

Le, X., Lancashire, I., Hirst, G., and 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):435–461.

Le Couteur, D. G., Doust, J., Creasey, H., and Brayne, C. (2013). Political drive to screen for pre-dementia: Not evidence based and ignores the harms of diagnosis. *BMJ Online*, 347:1–6.

Le Dorze, G. and Bédard, C. (1998). Effects of age and education on the lexico-semantic content of connected speech in adults. *Journal of Communication Disorders*, 31(1):53–71.

Lee, J. Y. and Hahn, M. (2010). Automatic assessment of pathological voice quality using higher-order statistics in the LPC residual domain. *EURASIP Journal on Advances in Signal Processing*, 2009(1):1–8.

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning*, 50(4):675–724.

Lehr, M., Prud'hommeaux, E. T., Shafran, I., and Roark, B. (2012). Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Proceedings of INTERSPEECH*, pages 1039–1042.

Lehr, M., Shafran, I., Prud'hommeaux, E. T., and Roark, B. (2013). Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 211–220.

Leyton, C. E. and Hodges, J. R. (2014). Differential diagnosis of primary progressive aphasia variants using the international criteria. *Aphasiology*, 28(8):909–921.

Li, H. (2011). Partial least squares-discriminant analysis and variable selection for high dimensional data (Matlab package).

Lin, F. and Weng, F. (2008). Computing confidence scores for all sub parse trees. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 217–220.

Little, M., McSharry, P., Moroz, I., and Roberts, S. (2006). Nonlinear, biophysically-informed speech pathology detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1080–1083.

Liu, Y. and Shriberg, E. (2007). Comparing evaluation metrics for sentence boundary detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 182–185. IEEE.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1526–1540.

Liu, Y., Stolcke, A., Shriberg, E., and Harper, M. (2005). Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 451–458.

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

Lubetich, S. and Sagae, K. (2014). Data-driven measurement of child language development with simple syntactic templates. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 2151–2160.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Lawrence Erlbaum Associates, Mahwah, New Jersey, 3 edition.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Manning, C. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? *Computational Linguistics and Intelligent Text Processing*, pages 171–189.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Martin, S., Kelly, S., Khan, A., Cullum, S., Dening, T., Rait, G., Fox, C., Katona, C., Cosco, T., Brayne, C., and Lafortune, L. (2015). Attitudes and preferences towards screening for dementia: A systematic review of the literature. *BMC Geriatrics*, 15(1):66–78.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, 23(2):143–149.

Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in Alzheimer's Disease: Can temporal and acoustic parameters discriminate dementia. *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

Mengistu, K., Rudzicz, F., and Falk, T. (2011). Using acoustic measures to predict automatic speech recognition performance for dysarthric speakers. In *Proceedings of the 7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*.

Mengistu, K. T. and Rudzicz, F. (2011). Adapting acoustic and lexical models to dysarthric speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4924–4927.

Menn, L. (1995). *Non-fluent Aphasia in a Multilingual World.* John Benjamins Publishing.

Messner, D. A. (2011). Informed choice in direct-to-consumer genetic testing for Alzheimer and other diseases: Lessons from two cases. *New Genetics and Society*, 30(1):59–72.

Mesulam, M.-M., Wieneke, C., Rogalski, E., Cobia, D., Thompson, C., and Weintraub, S. (2009). Quantitative template for subtyping primary progressive aphasia. *Archives of Neurology*, 66(12):1545–1551.

Mesulam, M.-M., Wieneke, C., Thompson, C., Rogalski, E., and Weintraub, S. (2012). Quantitative classification of primary progressive aphasia at early and mild impairment stages. *Brain*, 135(5):1537–1553.

Meteyard, L. and Patterson, K. (2009). The relation between content and structure in language production: An analysis of speech errors in semantic dementia. *Brain and Language*, 110(3):121–134.

Meteyard, L., Quain, E., and Patterson, K. (2014). Ever decreasing circles: Speech production in semantic dementia. *Cortex*, 55:17–29.

Miller, J. and Weinert, R. (1998). *Spontaneous Spoken Language*. Clarendon Press.

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., and Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of Neurology*, 49(12):1253–1258.

Munoz, D. G., Woulfe, J., and Kertesz, A. (2007). Argyrophilic thorny astrocyte clusters in association with Alzheimer's disease pathology in possible primary progressive aphasia. *Acta Neuropathologica*, 114(4):347–357.

Nadeau, S. E. (2012). *The Neural Architecture of Grammar*. MIT Press.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems (NIPS)*, volume 2, pages 841–848.

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

Okazaki, N. (2007). CRFsuite: A fast implementation of conditional random fields (CRFs).

Oppenheim, G. (1994). The earliest signs of Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology*, 7(2):116–120.

Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 78–87.

Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tür, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J., Liu, Y., Maskey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W., and Woofers, C. (2008). Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, 25(3):59 –69.

Paetzold, G. H. and Specia, L. (2016). Inferring psycholinguistic properties of words. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 435–440.

Pakhomov, S. V., Marino, S. E., and Birnbaum, A. K. (2013). Quantification of speech disfluency as a marker of medication-induced cognitive impairment: An application of computerized speech analysis in neuropharmacology. *Computer Speech & Language*, 27(1):116–134.

Pakhomov, S. V., Smith, G. E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., and Knopman, D. S. (2010a). Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23:165–177.

Pakhomov, S. V., Smith, G. E., Marino, S., Birnbaum, A., Graff-Radford, N., Caselli, R., Boeve, B., and Knopman, D. D. (2010b). A computerized technique to asses language use patterns in patients with frontotemporal dementia. *Journal of Neurolinguistics*, 23:127–144.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222.

Parsey, C. M. and Schmitter-Edgecombe, M. (2013). Applications of technology in neuropsychological assessment. *The Clinical Neuropsychologist*, 27(8):1328–1361.

Patterson, K., Graham, N., Lambon Ralph, M. A., and Hodges, J. (2006). Progressive nonfluent aphasia is not a progressive form of non-fluent (post-stroke) aphasia. *Aphasiology*, 20(9):1018–1034.

Patterson, K. and MacDonald, M. C. (2006). Sweet nothings: Narrative speech in semantic dementia. In Andrews, S., editor, *From Inkmarks to Ideas: Current Issues in Lexical Processing*. Psychology Press, Hove, UK.

Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (2008). Learning diagnostic models using speech and language measures. In *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4648–4651.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., and Ferri, C. P. (2013). The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's & Dementia*, 9(1):63–75.

Prins, R. and Bastiaanse, R. (2004). Review: Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, 18(12):1075–1091.

Prud'hommeaux, E., Morley, E., Rouhizadeh, M., Silverman, L., van Santen, J., Roark, B., Sproat, R., Kauper, S., and DeLaHunta, R. (2014). Computational analysis of trajectories of linguistic development in autism. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 266–271. IEEE.

Prud'hommeaux, E. and Roark, B. (2015). Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578.

Prud'hommeaux, E. T. (2012). *Alignment of narrative retellings for automated neuropsychological assessment*. PhD thesis, Oregon Health and Science University.

Prud'hommeaux, E. T. and Roark, B. (2011). Alignment of spoken narratives for automated neuropsychological assessment. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 484–489.

Prud'hommeaux, E. T. and Roark, B. (2012). Graph-based alignment of narratives for automated neurological assessment. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 1–10. Association for Computational Linguistics.

Prud'hommeaux, E. T., Roark, B., Black, L. M., and van Santen, J. (2011). Classification of atypical language in autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, CMCL '11, pages 88–96.

Quatieri, T. F. (2002). *Discrete-time Speech Signal Processing: Principles and Practice*. Prentice Hall.

Rapcan, V., D'Arcy, S., Penard, N., Robertson, I. H., and Reilly, R. B. (2009). The use of telephone speech recordings for assessment and monitoring of cognitive function in elderly people. In *Proceedings of INTERSPEECH*, pages 943–946.

Rapp, A. M. and Wild, B. (2011). Nonliteral language in Alzheimer dementia: A review. *Journal of the International Neuropsychological Society*, 17(02):207–218.

Raven, J. C. (1962). *Coloured Progressive Matrices Sets A, AB, B*. H. K. Lewis, London.

Read, J., Dridan, R., Oepen, S., and Solberg, L. J. (2012). Sentence boundary detection: A long solved problem? In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 985–994.

Reed, V. A., MacMillan, V., and McLeod, S. (2001). Elucidating the effects of different definitions of 'utterance' on selected syntactic measures of older children's language samples. *Asia Pacific Journal of Speech, Language and Hearing*, 6(1):39–45.

Reilly, J., Troche, J., and Grossman, M. (2011). Language processing in dementia. In Budson, A. E. and Kowall, N. W., editors, *The Handbook of Alzheimer's Disease and Other Dementias*, pages 336–368. Wiley-Blackwell.

Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C. A., and Garrard, P. (2014). Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease*, 42(S3):S3–S17.

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives de Psychologie*, 28:286–340.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090.

Robillard, J. M. (2016). The online environment: A key variable in the ethical response to complementary and alternative medicine for Alzheimer's disease. *Journal of Alzheimer's Disease*, 51(1):11–13.

Robillard, J. M., Illes, J., Arcand, M., Beattie, B. L., Hayden, S., Lawrence, P., McGrenere, J., Reiner, P. B., Wittenberg, D., and Jacova, C. (2015). Scientific and ethical features of English-language online tests for Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(3):281–288.

Robinson, K. M., Grossman, M., White-Devine, T., and D'Esposito, M. (1996). Category-specific difficulty naming with verbs in Alzheimer's disease. *Neurology*, 47(1):178–182.

Rochon, E., Saffran, E. M., Berndt, R. S., and Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language*, 72(3):193–218.

Rogalski, E., Cobia, D., Harrison, T. M., Wieneke, C., Thompson, C. K., Weintraub, S., and Mesulam, M.-M. (2011). Anatomy of language impairments in primary progressive aphasia. *Journal of Neuroscience*, 31(9):3344–3350.

Rohrer, J. D., Knight, W. D., Warren, J. E., Fox, N. C., Rossor, M. N., and Warren, J. D. (2008). Word-finding difficulty: A clinical analysis of the progressive aphasias. *Brain*, 131:8–38.

Rüping, S. (2006). *Learning Interpretable Models*. PhD thesis, Dortmund University.

Saffran, E. M., Berndt, R. S., and Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, 37(3):440–479.

Sagae, K., Lavie, A., and MacWhinney, B. (2005). Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 197–204.

Sajjadi, S. A., Patterson, K., Tomek, M., and Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, 26(6):847–866.

Salmon, D. P., Butters, N., and Chan, A. S. (1999). The deterioration of semantic memory in Alzheimer's disease. *Canadian Journal of Experimental Psychology*, 53(1):108–117.

Sampson, G. (1997). Depth in English grammar. *Journal of Linguistics*, 33:131–51.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). Technical Report MS-CIS-90-47, University of Pennsylvania.

Schatzmann, J., Thomson, B., and Young, S. (2007). Error simulation for training statistical dialogue systems. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 526–531.

Schicktanz, S., Schweda, M., Ballenger, J. F., Fox, P. J., Halpern, J., Kramer, J. H., Micco, G., Post, S. G., Thompson, C., Knight, R. T., and Jagust, W. J. (2014). Before it is too late: Professional responsibilities in late-onset Alzheimer's research and pre-symptomatic prediction. *Frontiers in Human Neuroscience*, 8:1–6.

Seidl, U., Lueken, U., Thomann, P. A., Kruse, A., and Schröder, J. (2012). Facial expression in Alzheimer's disease: Impact of cognitive deficits and neuropsychiatric symptoms. *American Journal of Alzheimer's Disease and Other Dementias*, 27(2):100–106.

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. In *Proceedings of INTERSPEECH*, volume 5, pages 1781–1784.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154.

Silva, D. G., Oliveira, L. C., and Andrea, M. (2009). Jitter estimation algorithms for detection of pathological voices. *EURASIP Journal on Advances in Signal Processing*, 2009:1–9.

Simons, J. S., Graham, K. S., Galton, C. J., Patterson, K., and Hodges, J. R. (2001). Semantic knowledge and episodic memory for faces in semantic dementia. *Neuropsychology*, 15(1):101–114.

Snowdon, David A .and Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Journal of the American Medical Association*, 275(7):528–532.

Solomon, P. R. and Murphy, C. A. (2005). Should we screen for Alzheimer's disease? *Geriatrics*, 60(11):26–31.

Solorio, T. (2013). Survey on emerging research on the use of natural language processing in clinical language assessment of children. *Language and Linguistics Compass*, 7(12):633–646.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Stadthagen-Gonzalez, H. and Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38(4):598–605.

Stevenson, M. and Gaizauskas, R. (2000). Experiments on sentence boundary detection. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 84–89.

Stolcke, A. and Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 1005–1008.

Strassel, S. (2005). *Topic Detection and Tracking Annotation Guidelines: Task Definition to Support the TDT2002 and TDT2003 Evaluations in English, Chinese and Arabic*. Linguistic Data Consortium, 1.5 edition.

Stromswold, K., Caplan, D., Alpert, N., and Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language*, 52(3):452–473.

Swanson, B. and Charniak, E. (2012). Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 193–197.

Taler, V. and Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Psychology*, 30(5):501–556.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). *Speech Coding and Synthesis*, 495:495–518.

Tanaka-Ishii, K. and Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica*, 65(1):96–116.

Tetreault, J., Foster, J., and Chodorow, M. (2010). Using parse features for preposition selection and error detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 353–358.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Proceedings of the IEEE International Conference on Mechatronics and Automation*, pages 1569–1574.

Thompson, C., Ballard, K., Tait, M., Weintraub, S., and Mesulam, M. (1997a). Patterns of language decline in non-fluent primary progressive aphasia. *Aphasiology*, 11:297–321.

Thompson, C., L., S., Tait, M., Jacobs, B., Schneider, S., and Ballard, K. (1995). A system for the linguistic analysis of agrammatic language production. *Brain and Language*, 51(1):124–129.

Thompson, C. K. and Bastiaanse, R. (2012). Introduction to agrammatism. In Bastiaanse, R. and Thompson, C. K., editors, *Perspectives on Agrammatism*, chapter 1, pages 1–16. Psychology Press, New York.

Thompson, C. K., Cho, S., Hsu, C.-J., Wieneke, C., Rademaker, A., Weitner, B. B., Mesulam, M. M., and Weintraub, S. (2012). Dissociations between fluency and agrammatism in primary progressive aphasia. *Aphasiology*, 26(1):20–43.

Thompson, C. K., Lange, K., Schneider, S. L., and Shapiro, L. P. (1997b). Agrammatic and non-brain-damaged subjects' verb and verb argument structure production. *Aphasiology*, 11(4-5):473–490.

Thompson, C. K. and Mack, J. E. (2014). Grammatical impairments in PPA. *Aphasiology*, 28(8):1018–1037.

Thompson, C. K., Meltzer-Asscher, A., Cho, S., Lee, J., Wieneke, C., Weintraub, S., and Mesulam, M. M. (2013). Syntactic and morphosyntactic processing in stroke-induced and primary progressive aphasia. *Behavioural Neurology*, 26(1):35–54.

Tierney, M. C. and Lermer, M. A. (2009). Computerized cognitive assessment in primary care to identify patients with suspected cognitive impairment. *Journal of Alzheimer's Disease*, 20(3):823–832.

Tomoeda, C. K., Bayles, K. A., Trosset, M. W., Azuma, T., and McGeagh, A. (1996). Cross-sectional analysis of Alzheimer disease effects on oral discourse in a picture description task. *Alzheimer Disease & Associated Disorders*, 10(4):204–215.

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., and Szatlóczki, G. (2015). Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In *Proceedings of INTERSPEECH*, pages 2694–2698. ISCA.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 252–259.

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., and Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271.

Van Velzen, M. and Garrard, P. (2008). From hindsight to insight: Retrospective analysis of language written by a renowned Alzheimer's patient. *Interdisciplinary Science Reviews*, 33(4):278–286.

Varley, R. (1993). Deictic terms, lexical retrieval and utterance length in aphasia: An investigation of inter-relations. *International Journal of Language & Communication Disorders*, 28(1):23–41.

Vertanen, K. (2006). Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments. Technical report, Cavendish Laboratory, University of Cambridge.

Vipperla, R., Renals, S., and Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices. In *Proceedings of INTERSPEECH*, pages 2550–2553.

Voss, M. J. (2005). Determining syntactic complexity using very shallow parsing. Master's thesis, University of Georgia.

Wang, D., Lu, L., and Zhang, H.-J. (2003). Speech segmentation without speech recognition. In *Proceedings of the IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 468–471. IEEE.

Warrington, E. K. and James, M. (1991). *The Visual Object and Space Perception Battery*. Thames Valley Test Company, Bury St Edmunds.

Webster, J., Franklin, S., and Howard, D. (2007). An analysis of thematic and phrasal structure in people with aphasia: What more can we learn from the story of Cinderella? *Journal of Neurolinguistics*, 20(5):363–394.

Webster, J. and Howard, D. (2012). Assessment of agrammatic language. In Bastiaanse, R. and Thompson, C. K., editors, *Perspectives on Agrammatism*, chapter 9, pages 136–157. Psychology Press, New York.

Weiner, M. F., Neubecker, K. E., Bret, M. E., and Hynan, L. S. (2008). Language in Alzheimer's disease. *The Journal of Clinical Psychiatry*, 69(8):1223–1227.

Wilpon, J. G. and Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 349–352. IEEE.

Wilson, J. M. G. and Jungner, G. (1968). Principles and practice of screening for disease. In *Public Health Papers*, number 34. World Health Organization.

Wilson, S. M., Galantucci, S., Tartaglia, M. C., and Gorno-Tempini, M. L. (2012). The neural basis of syntactic deficits in primary progressive aphasia. *Brain and Language*, 122(3):190–198.

Wilson, S. M., Galantucci, S., Tartaglia, M. C., Rising, K., Patterson, D. K., Henry, M. L., Ogar, J. M., DeLeon, J., Miller, B. L., and Gorno-Tempini, M. L. (2011). Syntactic processing depends on dorsal language tracts. *Neuron*, 72(2):397–403.

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., Miller, B. L., and Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain*, 133:2069–2088.

Wong, S.-M. J. and Dras, M. (2010). Parser features for sentence grammaticality classification. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 67–75.

Wong, S.-M. J. and Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1600–1610.

Woollams, A. M., Cooper-Pye, E., Hodges, J. R., and Patterson, K. (2008). Anomia: A doubly typical signature of semantic dementia. *Neuropsychologia*, 46(10):2503–2514.

Yancheva, M., Fraser, K., and Rudzicz, F. (2015). Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of* *the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 134–140.

Yngve, V. (1960). A model and hypothesis for language structure. *Proceedings of the American Physical Society*, 104:444–466.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1997). *The HTK book*. Cambridge University Engineering Department.

Young, V. and Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112.

Yunusova, Y., Graham, N. L., Shellikeri, S., Phuong, K., Kulkarni, M., Rochon, E., Tang-Wai, D. F., Chow, T. W., Black, S. E., Zinman, L. H., and Green, J. R. (2016). Profiling speech and pausing in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). *PloS One*, 11(1):1–16.

Zechner, K., Higgins, D., Xi, X., and Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.

Zhou, L., Fraser, K. C., and Rudzicz, F. (2016). Speech recognition in Alzheimer's disease and in its assessment. Submitted.