# Spoken Language Derived Measures for Detecting Mild Cognitive Impairment

Brian Roark, *Member, IEEE*, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye

***Abstract*—Spoken responses produced by subjects during neuropsychological exams can provide diagnostic markers beyond exam performance. In particular, characteristics of the spoken language itself can discriminate between subject groups. We present results on the utility of such markers in discriminating between healthy elderly subjects and subjects with mild cognitive impairment (MCI). Given the audio and transcript of a spoken narrative recall task, a range of markers are automatically derived. These markers include speech features such as pause frequency and duration, and many linguistic complexity measures. We examine measures calculated from manually annotated time alignments (of the transcript with the audio) and syntactic parse trees, as well as the same measures calculated from automatic (forced) time alignments and automatic parses. We show statistically significant differences between clinical subject groups for a number of measures. These differences are largely preserved with automation. We then present classification results, and demonstrate a statistically significant improvement in the area under the ROC curve (AUC) when using automatic spoken language derived features in addition to the neuropsychological test scores. Our results indicate that using multiple, complementary measures can aid in automatic detection of MCI.**

***Index Terms*—Forced alignment, linguistic complexity, mild cognitive impairment (MCI), parsing, spoken language understanding.**

## I. Introduction

NATURAL language processing (NLP) techniques are often applied to electronic health records and other clinical datasets to assist in extracting information from unstructured text. Recent research has examined a potential alternative clinical use of NLP for processing patient language samples to assess language development [3], [4] or the impact of neurodegenerative impairments on speech and language [1]. In this paper, we present methods for automatically measuring speech characteristics and linguistic complexity of spoken language samples elicited during neuropsychological exams of elderly subjects, and examine the utility of these measures for discriminating between clinically defined groups.

Among the most important clinical research advances in brain aging in recent years has been a sharpening of focus to the earliest clinically detectable phases of incipient dementia, commonly known as mild cognitive impairment (MCI) [5]. It has been rapidly appreciated that "MCI" is a clinical syndrome that is heterogeneous in its presentation [6] as well as in its outcomes [7], [8]. It has been shown that some people will reliably progress from MCI to dementia while others remain stable for years [8], [9]. A minority return to normal cognitive function. The degree to which these variations are the result of differences in definitions (e.g., single domain versus multi-domain MCI), populations or assessment methods are active and critical areas of investigation [10]. They speak to the need for multiple standardized approaches to these key studies. For the current study, we use the Clinical Dementia Rating (CDR [11], defined in Section II-A) to identify individuals with MCI. The CDR is assigned independently of the psychometric battery of exams investigated in this paper, and hence (importantly) provides an externally defined reference objective for evaluation.

MCI often goes undiagnosed. Clinicians in typical primary care practices find recognizing cognitive impairment at any stage challenging, failing to recognize or evaluate up to 50% of even later stage dementia [12]. Further, widely used screening tests such as the Mini-Mental State Examination (MMSE) do not reliably detect the relatively subtle impairment present in early MCI. Linguistic memory tests, such as word list and narrative recall, are more effective than the MMSE in detecting MCI, yet are still individually insufficient for adequate discrimination between healthy and impaired subjects. Because of this, a battery of examinations is typically used to improve psychometric classification. Yet the summary recall scores derived from these linguistic memory tests (total story units correctly recalled) disregard potentially useful information in the characteristics of the spoken language itself.

Previous work on the relationship between Alzheimer's disease and linguistic complexity demonstrates that low linguistic ability in early life has a strong association with cognitive impairment and Alzheimer's disease in later life [13]. In this paper, we examine a number of linguistic complexity measures [14], [15] and pause statistics [16] to find early markers of Alzheimer's in later life. Measures derived from characteristics of the spoken language go beyond simply measuring fidelity

Manuscript received August 06, 2010; revised November 05, 2010; accepted January 17, 2011. Date of publication February 07, 2011; date of current version nulldate. This research was supported in part by the National Science Foundation (NSF) under Grants #IIS-0447214 and #BCS-0826654, National Institute of Health/National Institute of Aging (NIH/NIA) grants #P30AG08017 and #R01AG024059, and pilot grants from the Oregon Center for Aging and Technology (ORCATECH, NIH #1P30AG024978-01) and the Oregon Partnership for Alzheimer's Research. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF or NIH. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

B. Roark and J.-P. Hosom are with the Center for Spoken Language Understanding, Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR 97239 USA (e-mail: roark@cslu.ogi.edu; hosom@cslu.ogi.edu).

M. Mitchell is with the Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, U.K. (e-mail: m.mitchell@abdn.ac.uk).

K. Hollingshead was with the Center for Spoken Language Understanding, Oregon Health and Science University, Portland, OR 97239 USA. She is now with the Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA (e-mail: hollingk@gmail.com).

J. Kaye is with the Departments of Neurology and Biomedical Engineering, and Layton Aging and Alzheimer's Disease Center, Oregon Health and Science University, Portland, OR 97239 USA, and also with the Portland Veterans Affairs Medical Center, Portland, OR, USA (e-mail: kaye@ohsu.edu).

Digital Object Identifier 10.1109/TASL.2011.2112351

1558-7916/$26.00 © 2011 IEEE

to the narrative, thus providing key additional dimensions for improved diagnosis of impairment. These may be analyzed for many characteristics of speech and language that the papers cited above have shown to discriminate between healthy subjects and subjects with impairment.

The key contributions of this work are threefold. First, we examine a large set of spoken language measures derived from clinically elicited samples, and demonstrate that they can be useful for discriminating between healthy and MCI groups. Second, we demonstrate the utility of these measures for a much older group of subjects than has been shown in the previous literature, with average age near 90. Finally, we demonstrate that effective automation of measure extraction is possible when given transcript and audio, so that significant differences in feature means between healthy and MCI groups are preserved. These automatically derived features are shown to provide statistically significant improvements in classification between healthy and MCI groups over the test scores alone. This paper extends results from two earlier conference publications [1], [2] by including a substantially larger set of subjects in the study—large enough to perform full classification trials—and by expanding the set of features that we investigate for this application. Overall, the results presented here are very promising for the use of automatically derived spoken language measures to augment the diagnostic utility of linguistic memory tests.

Working with samples taken from neuropsychological tests has several key benefits. First, methods will have direct clinical applicability, because they will apply to standard tests that are already in use in clinical settings. Second, to the extent that additional discriminative utility can be derived from the output of any particular neuropsychological test, the number of tests required for reliable screening will be reduced, leading to a better chance of widespread testing due to reduced demands on both clinicians and patients. Finally, the relatively constrained elicitation, focused on a fixed narrative, makes automation of marker extraction particularly feasible. The narrow and topic-focused use of language allows automatic extraction to be much more accurate for a large volume of subjects.

Most previous work on spoken language samples with such a population has focused on spontaneous conversational speech [14]–[16], leading to questions about whether the same useful characteristics derived in the previous work would be present in the clinically elicited samples. The narrow topic focus and formal setting that lead us to expect easier automation may also lead us to expect less variation among subjects. As we shall demonstrate, however, the relative ease of automation does not come at the expense of the utility of the spoken language markers: there remain enough differences in the spoken language to provide markers of good discriminative utility.

Additionally, the above cited works have focused on elderly subjects, but with mean ages in the 60s to mid 70s, hence yielding little evidence of whether such spoken language measures will be of utility for subjects beyond the age of 80. With increases in life expectancy, this is a growing segment of the population, and one where the risk of MCI and dementia are at their highest. With substantial changes in speech and language associated with typical aging, it is a real question as to whether measures which have been shown to discriminate between healthy subjects and subjects with MCI or dementia will continue to discriminate at more advanced ages. As we show in this paper, differences in spoken language do provide markers of good discriminative utility for subjects well above the age of 80. Due to all of these considerations, we believe that use of spoken language markers to diagnose MCI is a compelling area for further research.

For many measures of speech characteristics and linguistic complexity, the nature of the annotation is critical—different conventions of structural or temporal annotation will likely yield different scores. We will thus spend the next section briefly detailing the annotation conventions that were used for this work. This is followed by a section describing measures to be derived from these annotations. Finally, we present empirical results on the samples of spoken narrative retellings.

## II. Data and Annotation Conventions

### A. Subjects

The subjects in this study came from existing community cohort studies of brain aging at the Layton Aging & Alzheimer's Disease Center, an NIA-funded Alzheimer's center for research at Oregon Health & Science University (OHSU). The Layton Center defines MCI in two ways: 1) via the Clinical Dementia Rating (CDR) scale [11] and 2) via a psychometrically driven concept of degraded performance on a large set of neuropsychological tests. This paper is investigating speech produced during the neuropsychological tests that play a role in the latter definition of MCI. Thus, to provide an independent unconfounded reference objective for evaluation, our reference classification is based on the CDR scale, defined next. The CDR has been shown to have high expert inter-annotator reliability [17], and, critically, is not deterministically derived from neuropsychological test results.

The CDR is assigned with access to information from clinical examinations, and is calculated by assigning scores for six cognitive categories: Memory; Orientation; Judgment and Problem Solving; Community Affairs; Home and Hobbies; and Personal Care. The score for each category is an assessment of impairment in that category, and can take one of five values: 0 (None); 0.5 (Questionable); 1 (Mild); 2 (Moderate); and 3 (Severe). The information that is used to provide the scores in each of these categories comes from clinical interviews with the subject and an informant, and a clinician-administered Neurobehavioral Cognitive Status Examination. To derive the CDR from these independently scored categories, the score $M$ of the Memory category is taken as primary. The CDR is $M$ if at least three of the remaining five categories also have a score of $M$. Otherwise, the CDR is calculated by looking at the scores of categories with score other than $M$. See [11] for specific details.

We collected audio recordings of 74 neuropsychological examinations administered at the Layton Center at OHSU. The two subject groups were: 1) those who were assigned a Clinical Dementia Rating (CDR) of 0 (healthy); and 2) those who were assigned a CDR of 0.5 (Mild Cognitive Impairment; MCI). 37 subjects were in the $\mathrm{CDR} = 0$ group (healthy); and 37 subjects are in the $\mathrm{CDR} = 0.5$ group (MCI).

### B. Neuropsychological Tests

Table I presents means and standard deviations for age, years of education and the manually calculated scores of a number

TABLE I
NEUROPSYCHOLOGICAL TEST RESULTS FOR SUBJECTS.
$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

| Test | | CDR = 0 | | CDR = 0.5 | | |
|---|---|---|---|---|---|---|
| | | (n=37) | | (n=37) | | |
| ID | Measure | M | SD | M | SD | $t(72)$ |
| | Age | 88.8 | 6.0 | 89.8 | 5.7 | $-0.77$ |
| | Education (Y) | 15.1 | 2.4 | 14.5 | 2.8 | 0.95 |
| 1 | MMSE | 28.2 | 1.7 | 26.4 | 2.4 | $3.64^{***}$ |
| 2 | Word List (A) | 19.5 | 4.2 | 15.8 | 3.6 | $4.04^{***}$ |
| 3 | Word List (R) | 6.4 | 2.1 | 4.1 | 2.1 | $4.65^{***}$ |
| 4 | Wechsler LM I | 15.9 | 4.5 | 11.8 | 4.5 | $3.96^{***}$ |
| 5 | Wechsler LM II | 14.4 | 4.8 | 10.1 | 4.8 | $3.51^{***}$ |
| 6 | Cat.Fluency (A) | 17.8 | 4.5 | 14.0 | 4.1 | $3.78^{***}$ |
| 7 | Cat.Fluency (V) | 12.4 | 4.0 | 9.8 | 3.4 | $2.97^{**}$ |
| 8 | Digits (F) | 6.8 | 1.2 | 6.3 | 1.2 | $1.83^{*}$ |
| 9 | Digits (B) | 4.8 | 1.1 | 4.3 | 1.0 | $2.07^{*}$ |

of standard neuropsychological tests that were administered during the recorded session. Again, these tests are not used to assign the CDR. These tests include: the Mini Mental State Examination (MMSE) [18]; the CERAD Word List Acquisition (A) and Recall (R) tests [19]; the Wechsler Logical Memory (LM) I (immediate) and II (delayed) narrative recall tests [20]; Category Fluency, Animals (A) and Vegetables (V); and Digit Span (WAIS-R) forward (F) and backward (B) [21]. Test IDs are given in the table for reference later in the paper.

The Wechsler Logical Memory I/II tests are the basis of our study on spoken language derived measures. The original narrative is a short, three-sentence story:

> Anna Thompson of South Boston, employed as a cook in a school cafeteria, reported at the police station that she had been held up on State Street the night before and robbed of fifty-six dollars. She had four small children, the rent was due, and they had not eaten for two days. The police, touched by the woman's story, took up a collection for her.

Subjects are asked to re-tell this story immediately after it is told to them (LM I), as well as after approximately 30 minutes of unrelated activities (LM II). Each retelling is scored by assigning a point for 25 possible "story units," as defined in the scoring guidelines of the test. For example, using the word "cook" earns a point for one story unit; also the words "Thompson," "Boston," "school," "cafeteria," and "police" each count as story units, along with many other words and phrases, covering the entire short narrative. A subject's score for both the immediate and delayed retelling is the sum total of recalled story units for that retelling, as tallied by the examiner.

We transcribed each retelling and manually time-aligned the transcripts with the audio (see Section II-E). We also manually annotated syntactic parse trees according to the Penn Treebank annotation guidelines (see Section II-D). Algorithms for calculating linguistic complexity measures from parse trees were written to accept either manually annotated trees or trees output from an automatic parser. Similarly, algorithms for calculating speech measures, such as pause frequency and duration, were written to accept either manual or automatic time alignments. This is useful for a comparison of feature extraction from automatic and manual annotations.

### C. Manual Transcriptions

Conversational speech and other types of spontaneous language do not tend to occur in clearly defined sentences. We followed the utterance segmentation rules outlined in the 2004 EARS Official Annotation Guidelines [22] for conversational speech, such as that found in the Switchboard corpus [23]. These guidelines differ substantially from the original Switchboard treebank guidelines from the 1999 release, which relied heavily on speech stream characteristics and imposed a 15-s cutoff for utterances. The 2004 guidelines were used for the most current LDC release of the Switchboard treebank. Briefly, utterances are separated based on whether 1) they contain a subject and a predicate, or 2) if they have no subject, they contain a predicate that is not coordinated with the previous predicate. The exception to this is embedded clauses, which were included within their matrix clauses.

In some cases, the speech was too fragmented to make such clear-cut distinctions. In these cases, we separated out each fragment if it did not clearly belong to the previous or following utterance. This was determined both by the semantic content of the utterances and whether there was a pause (>1 s) between the fragment and the surrounding speech. All transcription work was completed by one author, ensuring consistency across the different transcriptions.

For manual time-alignment of the transcript with the audio, we marked the start and end points of each word and pause in the speech stream. Manual transcriptions were also annotated to mark *reparandums*, a sequence of words that are aborted by the speaker then repaired within the same utterance; and *parentheticals*, an utterance within a narrative that is not part of the narrative itself. Parentheticals longer than one full utterance were considered to be outside of the narrative recall, and were not included in our analyses. We discuss these cases in further detail below. Additionally, we isolated only the narrative recall itself; conversational speech that occurred immediately before and immediately after the recall was disregarded.

### D. Syntactic Annotation

For manual syntactic annotation of collected data, we followed the syntactic annotation conventions of the Penn Treebank [24]. This provides several key benefits. First, there is an extensive annotation guide that has been developed, not just for written but also for spoken language, so that consistent annotation was facilitated. Second, large out-of-domain corpora, in particular the 1 million words of syntactically annotated Switchboard telephone conversations, provide a good starting point for training domain-adapted parsing models. Finally, we can use multiple domains for evaluating the correlations between various linguistic complexity measures.

There are characteristics of Penn Treebank annotation that can impact linguistic complexity scoring. First, prenominal modifiers are typically grouped in a flat constituent with no internal structure. This annotation choice can result in very long noun phrases (NPs) which pose very little difficulty in terms of human processing performance, but can inflate complexity measures that measure deviation from right-branching structures, such as that of Yngve [25], described in Section III-B. Second, in spoken language annotations, a reparandum is denoted with a special non-terminal category EDITED; and parentheticals are denoted with a special non-terminal category PRN. For this paper, these are treated as constituents.

### *E. Pause Annotations From Time Alignments*

Given time alignments marking the start and end points of each word and pause in the speech stream, we can calculate pause measures. Background noises and non-speech sounds (coughing, tongue-clicks, etc.), were not marked, and were generally subsumed within the pause time-alignments. Interjections, such as "um," "uh," and "hmm," were transcribed; however, these were analyzed as filled pauses and so added to the total pause count and pause time. All pauses in the narrative recall greater than one second were included within our pause statistics.

Crosstalk between the interviewer and the subject occurred rarely, but in these cases we transcribed the subject only. This was also best for recognition purposes, as the subject was loudest in the speech stream due to microphone placement. Speech produced by the interviewer alone was marked as such, and not included in our analyses. These interjections are minimal, making up 1.06% of the time in our data. They consist of sounds of agreement ("mm-hmm," "yeah") and a reprompt ("Are there any other details you can think of?").

## III. Spoken Language Derived Measures

### *A. Approaches to Linguistic Complexity*

There is no single agreed-upon measurement of linguistic complexity in the literature. A range of measures have been proposed, with different primary considerations driving the notion of complexity for each. Many measures focus on the order in which various constructions are acquired by children learning the syntax of their native language—later acquisitions being taken as higher complexity. Examples of this sort of complexity measure are: mean length of utterance (MLU), which is typically measured in morphemes [26]; the Index of Productive Syntax [27], a multi-point scale which has recently been automated for child-language transcript analysis [3]; and Developmental Level [28], a 7-point scale of complexity based on the presence of specific grammatical constructions.

For assessing adult language—the focus of the current study—approaches have relied upon the right-branching nature of English syntactic trees [25], [29], under the assumption that deviations from that correspond to more complexity in the language. Additionally, there are approaches focused on the memory demands imposed by "distance" between dependent words [30], [31]. We next discuss in detail the particular linguistic complexity measures we employed in the study presented here, followed by other spoken language derived measures.

### *B. Yngve Scoring*

The scoring approach taken in Yngve [25] is related to the size of a "first in/last out" (pushdown) stack at each word in a top-down, left-to-right parse derivation. To calculate the size of the stack at each word, we can use the following simple algorithm. At each node in the tree, score the branches from that node to each of its children, beginning with a score of zero at the rightmost child and continuing to the leftmost child, incrementing the score by one for each child. Hence, each rightmost branch in the tree of Fig. 1 (solid lines) is labeled with 0, the leftmost branch in all binary nodes is labeled with 1, and the leftmost branch in the ternary node is labeled with 2. Then the score for each word is the sum of the branch scores from the root of the tree to the word.

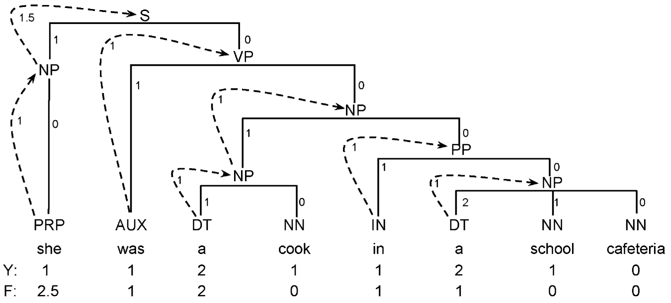
S
NP
VP
NP
PP
NP
NP
1.5
PRP AUX DT NN IN DT NN NN
she was a cook in a school cafeteria
Y: 1 1 2 1 1 2 1 0
F: 2.5 1 2 0 1 1 0 0

Fig. 1. Parse tree with branch scores for Yngve scoring, and upward link scores (on dotted lines) for Frazier scoring. The Yngve score (Y) and Frazier score (F) is shown for each word.

Given the score for each word, we can then derive an overall complexity score by summing them or taking the maximum or mean. For this paper, we report mean scores for this and other word-based measures, since we have found these means to provide better performing scores than either total sum or maximum. For the tree in Fig. 1, the maximum is 2, the total is 9 and the mean over 8 words is $1(1/8)$.

### *C. Frazier Scoring*

Frazier [29] proposed an approach to scoring linguistic complexity that traces a path from a word up the tree until reaching either the root of the tree or the lowest node that is not the leftmost child of its parent (i.e., finding a sibling to its left).[1] For example, the tree in Fig. 1 is augmented with dotted arcs that trace a path from words up to the node specified by Frazier's approach. The first word has a path up to the root S node, while the second word just up to the VP, since the VP has an NP sibling to its left. The word is then scored, as in the Yngve measure, by summing the scores on the links along the path. Each non-terminal node in the path contributes a score of 1, except for sentence nodes and sentence-complement nodes,[2] which score 1.5 rather than 1. Thus, embedded clauses contribute more to the complexity measure than other embedded categories, as an explicit acknowledgment of sentence embeddings as a source of linguistic complexity.

As with the Yngve score, we can calculate the total and the mean of these word scores. For the example in Fig. 1, the maximum is 2.5 and the total is 7.5, yielding a mean of $(15/16)$.

### *D. Dependency Distance*

Rather than examining the tree structure itself, one might also extract measures from lexical dependency structures. These dependencies can be derived from the tree using standard rules for establishing head children for constituents, originally attributed to Magerman [32], to percolate lexical heads up the tree. Fig. 2 shows the dependency graph that results from this head percolation approach, where each link in the graph represents a dependency relation from the modifier to the head. For example,

[1] An exception is made for empty subject NPs, in which case the succeeding verb receives an additional score of 1 (for the deleted NP), and its path continues up the tree.

[2] Every non-terminal node beginning with an S, including SQ and SINV, were counted as sentence nodes. Sequences of sentence nodes, i.e., an SBAR appearing directly under an S node, were only counted as a single sentence node and thus only contributed to the score once.

conventional head percolation rules specify the VP as the head of the S, so "was," as the head of the VP, is thus the lexical head of the entire sentence. The lexical heads of the other children of the S node are called modifiers of the head of the S node; thus, since "she" is the head of the subject NP, there is a dependency relation between "she" and "was."

Lin [30] argued for the use of this sort of dependency structure to measure the difficulty in processing, given the memory overhead of very long distance dependencies. Both Lin [30] and Gibson [31] showed that human performance on sentence processing tasks could be predicted with measures of this sort. While details may differ—e.g., how to measure distance, what counts as a dependency—we can make use of the general approach given Treebank style parses and head percolation, resulting in graphs of the sort in Fig. 2. For the current paper, we count the distance between words for each dependency link. For Fig. 2, there are 7 dependency links, a distance total of 11, and a mean of $1(4/7)$.

### E. POS-Tag Sequence Cross Entropy

One possible approach for detecting rich syntactic structure is to look for infrequent or surprising combinations of parts-of-speech (POS). We can measure this over an utterance by building a simple bi-gram model over POS tags, then measuring the cross entropy of each utterance.[3]

Given a bi-gram model over POS-tags, we can calculate the probability of the sequence as a whole. Let $\tau_i$ be the POS-tag of word $w_i$ in a sequence of words $w_1 \ldots w_n$, and assume that $\tau_0$ is a special start symbol, and that $\tau_{n+1}$ is a special stop symbol. Then the probability of the POS-tag sequence is

$$\mathrm{P}(\tau_1 \ldots \tau_n) = \prod_{i=1}^{n+1} \mathrm{P}(\tau_i \mid \tau_{i-1}). \tag{1}$$

The cross entropy is then calculated as

$$H(\tau_1 \ldots \tau_n) = -\frac{1}{n} \log \mathrm{P}(\tau_1 \ldots \tau_n). \tag{2}$$

With this formulation, this boils down to the mean negative log probability of each tag given the previous tag. For this study, we derive POS-tags from the output of parsing.

We examine POS-tag cross entropy in two conditions, calculated using models derived just from the Switchboard treebank as well as models adapted to this domain. The details of domain adaptation are discussed in further detail in Section IV-A.

### F. Alternative Tree Analyses

For the tree-based complexity metrics (Frazier and Yngve), we also investigated alternative implementations that make use of systematic transformations of the tree structure. None of these yielded much difference in our analysis versus the standard trees, but we mention them here for completeness.

One alternative was the left-corner transformation [33] of the tree from which the measures were calculated. In Roark *et al.* [2], we presented results using either manually annotated trees or automatic parses to calculate the Yngve and Frazier measures after a left-corner transform had been applied to the tree. The patterns were very similar to those achieved without the transform; hence, the results are omitted in this paper.

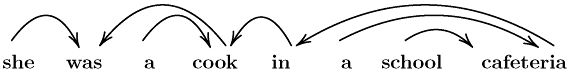

Fig. 2. Dependency graph for the example string.

We also calculated the Frazier and Yngve measures using only the embedded nodes within each tree, as a way to examine the effects of center-embedding, but this did not yield measures that showed statistically significant group differences; hence, the raw results are omitted in this paper.

Finally, we looked at the effects of removing disfluencies (EDITED nodes in the trees) and parentheticals (PRN). These did yield some interesting differences from the original trials. In particular, when removing these constituents, words per retelling shows statistically significant group differences in both the immediate and delayed retelling conditions, which is not the case when they are left in. Also, POS-tag cross entropy shows statistically significant group differences in all conditions using either domain-adapted models or non-adapted models. Only domain-adapted models yield statistically significant group differences using the standard structures. There are other minor differences, but not enough to warrant a detailed examination of this transform in this paper.

### G. Idea Density

Another way to measure the complexity of a sentence using POS-tags is to look at the number of propositions expressed in the sentence. This is the metric of idea density (also known as propositional density or P-density) [13], which is defined as the number of expressed propositions divided by the number of words. This provides a way to measure how many assertions a speaker is making. In this metric, propositions correspond to verbs, adjectives, adverbs, prepositions, and conjunctions. Nouns are not considered to be propositions, as the main verb and all its arguments count as one proposition. Calculation of P-density was performed using a modified version of the rules outlined in the Computerized Propositional Idea Density Rater (CPIDR) program [34], adapted to the POS-tags used in the Penn Treebank.

### H. Content Density

A related measure that we explored is the ratio of open-class words to closed-class words, which we term *content density*. An open word class is one in which an unlimited number of new words may be added and created, such as nouns and verbs; while closed word classes are fixed (relatively small) sets, such as prepositions. This is also stated as the distinction between content words and function words.

This metric is comparable to the P-density (idea density) metric, except that nouns are counted in addition to verbs. Content density is calculated over POS-tags, where open-class words are those with POS-tags corresponding to nouns, verbs, adjectives, adverbs, and symbols, i.e., those POS-tags beginning with NN, VB, JJ, RB, or SYM. The rest are considered closed-class words.[4]

[3] For each test domain, we used cross-validation techniques to build POS-tag bi-gram models and evaluate with them in that domain.

[4] Interjections are technically an open class, but for our purposes, we group interjection tags with other non-content tags.

TABLE II
PARSER ACCURACY ON WECHSLER LOGICAL MEMORY RESPONSES USING JUST OUT-OF-DOMAIN DATA (SWITCHBOARD (SWBD) TREEBANK) VERSUS USING A DOMAIN-ADAPTED SYSTEM

| System | Labeled Recall | Labeled Precision | F-measure |
|---|---|---|---|
| Out-of-domain (SWBD) | 83.2 | 86.4 | 84.8 |
| Domain-adapted from SWBD | 87.7 | 89.0 | 88.3 |

TABLE III
CORRELATION MATRICES FOR SEVERAL MEASURES ON AN UTTERANCE-BY-UTTERANCE BASIS. CORRELATIONS ALONG THE DIAGONAL ARE BETWEEN THE MANUAL MEASURES AND THE MEASURES WHEN AUTOMATICALLY PARSED. ALL OTHER CORRELATIONS ARE BETWEEN MEASURES WHEN DERIVED FROM MANUAL PARSE TREES

| **Penn WSJ** | Frazier | Yngve | Tree nodes | Dep len |
|---|---|---|---|---|
| Frazier | 0.89 | | | |
| Yngve | -0.31 | 0.96 | | |
| Tree nodes | 0.91 | -0.16 | 0.92 | |
| Dependency length | -0.29 | 0.75 | -0.13 | 0.93 |
| **Penn SWBD** | | | | |
| Frazier | 0.96 | | | |
| Yngve | -0.72 | 0.96 | | |
| Tree nodes | 0.58 | -0.06 | 0.93 | |
| Dependency length | -0.74 | 0.97 | -0.08 | 0.96 |
| **Wechsler LM** | | | | |
| Frazier | 0.87 | | | |
| Yngve | -0.40 | 0.93 | | |
| Tree nodes | 0.99 | -0.43 | 0.87 | |
| Dependency length | -0.43 | 0.95 | -0.47 | 0.94 |

### I. *Speech Duration Measures*

In addition to measures derived from the syntactic structure of the utterances in the retelling, we derived measures from temporal aspects of the speech sample, such as number of pauses and total pause duration. We follow Singh *et al.* [16] in setting a 1-second minimum for counting silence as pause, and we also employ their definitions of the following speech-based measures. For these, words or phrases marked as EDITED (reparandum) or PRN (parenthetical comments) were included as standard words in the retelling. However, a set of pre-defined “filler” words were considered to be pause events instead of standard words. The list of filler words was {um, uh, er, eh, hm, huh, duh}. Speech measures included the following:

- The **Pauses per Retelling** measure is the total number of pauses per sample.
- The **Total Pause Time** is the total duration of all pauses (minimum 1-second duration) per sample, in seconds.
- The **Mean Duration of Pauses** is the Total Pause Time divided by the Pauses per Retelling.
- The **Standardized Pause Rate** is the number of words in the sample divided by the Pauses per Retelling.
- The **Total Phonation Time** is the amount of time, in seconds, in the sample that contains speech events.
- The **Total Locution Time** is the amount of time in the sample that contains both speech and pauses.
- The **Phonation Rate** is the Total Phonation Time divided by the Total Locution Time.
- The **Transformed Phonation Rate** is the arcsine of the square root of the Phonation Rate. According to Singh *et al.*, “This transformation provides a normally distributed measure within each participant group” [16].
- The **Standardized Phonation Time** is the number of words in the sample divided by the Total Phonation Time.
- The **Verbal Rate** is the number of words in the sample divided by the Total Locution Time.

## IV. EXPERIMENTAL RESULTS

### A. *Parsing*

For automatic parsing, we made use of the Charniak parser [35]. In this section, in order to evaluate parsing results under standard conditions, we present parse accuracy evaluation with the EDITED nodes removed [36], though later results in Section IV-D and IV-E do not remove these nodes. Table II shows parsing accuracy[5] of our annotated retellings under two parsing model training conditions: 1) trained on approximately 1 million words of Switchboard (SWBD) corpus telephone conversations; and 2) using domain adaptation techniques starting from the SWBD Treebank. The SWBD out-of-domain system reaches quite respectable accuracies, and domain adaptation achieves a 3.5% absolute improvement over that.

For domain adaptation, we used MAP adaptation techniques [38] via cross-validation over the entire set of retellings. For each subject, we trained a model using the SWBD treebank as the out-of-domain training data, and the retellings of the other 73 subjects as in-domain training. For our parser, we used a count merging approach, with the in-domain counts scaled by 1000 relative to the out-of-domain counts. See Bacchiani *et al.* [38] for more information on stochastic grammar adaptation using these techniques.

### B. *Correlations*

Our next set of experimental results present correlations between parse-tree derived measures. Table III shows results for four of our measures over both treebanks that we have been considering (Penn SWBD Treebank, and the Wechsler LM retellings) as well as for the Penn WSJ Treebank as a comparison point. The correlations along the diagonal are between the same measure when calculated from manually annotated trees and when calculated from automatic parses.[6] All other correlations are between measures derived from manual trees. All correlations are taken per utterance.

From this table, we can see that all of the measures derived from automatic parses have a high correlation with the manually derived measures, indicating that they may preserve any discriminative utility of these measures. Interestingly, simply calculating the number of nodes in the tree per word tends to correlate well with the Frazier score, while the dependency length

[5] We use the standard PARSEVAL measures [37] for evaluation of parse accuracy: labeled precision is the number of correct labeled spans divided by labeled spans in the automatic parse; labeled recall is the number of correct labeled spans divided by labeled spans in the true parse; and F-measure is the harmonic mean of these two scores.

[6] SWBD and WSJ automatic parses are trained from same-domain data, using standard training and development sets; for the Wechsler LM retellings, we used the parsing model adapted from out-of-domain SWBD data.

tends to correlate well with the Yngve score. These two groups of syntactic measures seem to roughly correspond to those measuring depth of the tree (Frazier, tree nodes) versus those measuring the breadth of the tree (Yngve, dependency length). Notably, there seems to be a slightly lower correlation between manual and automatic annotations for the depth measures than for the breadth measures.

### C. *Forced Alignment for Pause Durations*

Forced alignment uses an existing automatic speech recognition (ASR) system, and constrains it so that it can only recognize the (known) word sequence, with optional pauses between each word. The output then contains the location in the speech signal of each word and pause event. Our baseline forced alignment system [39] has performance comparable to, or better than, the best systems reported in the literature, placing 91.5% of phoneme boundaries within 20 ms of manual boundaries on the TIMIT corpus.[7] For the current task, this baseline forced-alignment system was re-trained with two primary modifications.

First, acoustic data from the current Wechsler LM corpus were used to train the system, in addition to data from several other corpora (TIMIT, OGI Stories, and OGI Portland Cellular), using a two-fold cross-validation procedure. Second, because of the different channel characteristics of the four corpora used in training the system, the standard feature set of 13 cepstral coefficients and their delta values was replaced by a set of spectral features. The spectral features included 26 values from the log power spectrum sampled along the Mel scale. Adaptive pre-emphasis was performed to remove the effects of spectral tilt, and this pre-emphasis value was included, along with total energy (in dB), for a total of 28 spectral features at each frame. The final feature set consisted of these 28 features and their delta values at the center frame, and these 28 features at frames $-60, -30$, 30, and 60 ms relative to the center frame, for a total of 168 feature values per frame.

Performance of the re-trained forced-alignment system was evaluated by measuring the difference in word boundaries obtained from manual annotation and from forced alignment. The baseline system from [39] placed 72.8% of word boundaries within 100 ms of manual boundaries on the MCI corpus. With the above domain adaptation techniques, the resulting system had average test-set performance of 84.0% of word boundaries placed within 100 ms of manual word boundaries on the MCI corpus. The remaining large error (16.0%) relative to other test sets indicates the sensitivity of even the re-trained system to background noise, including non-speech sounds and breath noise. The re-trained system tends to incorrectly assign small background noises, when surrounded by very long pauses, to speech events. While it is unknown to what degree these errors impact the significance of speech measures for MCI, improving performance of the forced-alignment system will be a subject of future research.

### D. *Group Differences*

Table IV presents means and standard deviations for measures derived from the LM I and LM II retellings, along with the $t$-value and level of significance. Feature IDs are given in the table for reference in the next section of the paper. The first three measures presented in the table are available without syntactic or time alignment annotation: total number of words, total number of utterances, and words per utterance in the retelling. None of these three measures on either retelling show statistically significant differences between the groups.

The first measure to rely upon syntactic annotations is words per clause.[8] The number of clauses are calculated from the parses by counting the number of S nodes in the tree.[9] Normalizing the number of words by the number of clauses rather than the number of utterances (as in words per utterance) results in statistically significant differences between the groups for both LM I and LM II. This pattern is evident in both the automated and the manual results.

The other language measures are as described in Section III. The Frazier score per word, the number of tree nodes per word, and the P-density score per word do not show a significant $t$-value on the LM retellings for either LM test.

The Yngve score per word and the dependency length per word show no significant difference on LM I retellings but a statistically significant difference on LM II. See Section V for a discussion of why differences should be expected between the immediate and delayed retellings.

The POS-tag cross entropy, when calculated using models adapted to this domain (Logical Memory tests)—denoted "POS cross entropy (LM)"—shows a statistically significant difference between groups for LM I. Without such domain adaptation—denoted "POS cross entropy (SW)" for Switchboard—we do not observe a statistically significant difference between groups. This is apparent in both the automatically and manually parsed/tagged data.

While P-Density did not show significant group differences, content density appears to be a particularly useful measure for discriminating between the two groups, with a statistically significant difference here for both LM tests. Given the relatively high quality of the automatic parses, most of the means and standard deviations are quite close, and almost all patterns derived from the manually parsed data are also true of the automatically parsed data.

For manual annotation, the Standardized Pause Rate for LM I and Total Phonation Time for LM I and LM II yield a significant difference between CDR 0 and CDR 0.5 subjects. For speech and pause annotation obtained from forced alignment, however, the Standardized Pause Rate, Phonation Rate, and Transformed Phonation Rate all show significant differences for LM I. Total Phonation Time is the only automatic measure that shows a significant difference for LM II.

It is not immediately clear why the forced-alignment system, which has only 84% agreement with manual annotation at large (100 ms) thresholds, yields significant differences for a larger number of speech measures than manual annotation. One possibility is that the long pauses that cause difficulty for the forced-alignment system are actually detrimental to the value of the measures, and "breaking up" these very long pauses, even if incorrectly, allows the measures to focus on the (hypothetically)

[7] Boundary agreement between two humans on the TIMIT corpus is 93.5% within 20 ms.

[8] Note that the automatic parsing used to derive the language measures did not remove EDITED or PRN substrings or nodes from the utterances or derived trees.

[9] For coordinated S nodes, the root of the coordination, which in Penn Treebank style annotation also has an S label, does not count as an additional clause.

TABLE IV
LANGUAGE AND SPEECH MEASURE GROUP DIFFERENCES WHEN MEASURES ARE DERIVED FROM EITHER MANUAL OR AUTOMATIC ANNOTATIONS. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

| | | Logical Memory I | | | | | Logical Memory II | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | | CDR=0 | | CDR=0.5 | | | CDR=0 | | CDR=0.5 | | |
| ID | Measure | **mean** | **SD** | **mean** | **SD** | *t(72)* | **mean** | **SD** | **mean** | **SD** | *t(72)* |
| 1 | **Words per retell** | 69.22 | 23.01 | 64.22 | 28.25 | 0.83 | 70.43 | 28.83 | 61.03 | 31.91 | 1.33 |
| 2 | **Utter. per retell** | 8.41 | 3.08 | 8.70 | 4.30 | -0.33 | 8.03 | 3.59 | 7.54 | 4.10 | 0.55 |
| 3 | **Words per utter.** | 8.72 | 2.60 | 7.91 | 3.57 | 1.12 | 9.28 | 3.01 | 8.09 | 4.14 | 1.41 |
| **Language Measures** | | | | | | | | | | | |
| 4 | **Words per clause** | 6.20 | 1.48 | 5.47 | 1.53 | 2.09* | 6.49 | 1.29 | 5.17 | 1.87 | 3.53*** |
| | Automated: | 6.23 | 1.58 | 5.25 | 1.34 | 2.88** | 6.34 | 1.29 | 5.15 | 1.79 | 3.28*** |
| 5 | **Frazier per word** | 1.07 | 0.10 | 1.10 | 0.11 | -1.23 | 1.05 | 0.07 | 1.05 | 0.27 | 0.00 |
| | Automated: | 1.05 | 0.08 | 1.08 | 0.09 | -1.52 | 1.03 | 0.08 | 1.02 | 0.26 | 0.22 |
| 6 | **Tree Nodes per word** | 2.00 | 0.09 | 2.03 | 0.10 | -1.36 | 1.98 | 0.07 | 1.93 | 0.47 | 0.64 |
| | Automated: | 1.98 | 0.08 | 2.01 | 0.09 | -1.52 | 1.96 | 0.07 | 1.90 | 0.47 | 0.77 |
| 7 | **Yngve per word** | 1.49 | 0.27 | 1.48 | 0.42 | 0.12 | 1.57 | 0.26 | 1.38 | 0.49 | 2.08* |
| | Automated: | 1.41 | 0.22 | 1.39 | 0.30 | 0.33 | 1.54 | 0.29 | 1.32 | 0.47 | 2.42** |
| 8 | **Depend Len per word** | 1.57 | 0.29 | 1.55 | 0.41 | 0.24 | 1.66 | 0.28 | 1.47 | 0.48 | 2.08* |
| | Automated: | 1.50 | 0.22 | 1.49 | 0.34 | 0.15 | 1.65 | 0.32 | 1.43 | 0.47 | 2.35* |
| 9 | **POS cross entropy (LM)** | 2.09 | 0.24 | 2.00 | 0.22 | 1.68* | 2.10 | 0.25 | 1.96 | 0.54 | 1.43 |
| | Automated: | 2.11 | 0.23 | 2.02 | 0.20 | 1.80* | 2.12 | 0.24 | 1.98 | 0.53 | 1.46 |
| 10 | **POS cross entropy (SW)** | 2.06 | 0.20 | 2.00 | 0.19 | 1.32 | 2.08 | 0.21 | 1.94 | 0.51 | 1.54 |
| | Automated: | 2.07 | 0.20 | 2.01 | 0.17 | 1.39 | 2.08 | 0.20 | 1.94 | 0.51 | 1.55 |
| 11 | **P-Density per word** | 0.40 | 0.07 | 0.41 | 0.09 | -0.53 | 0.44 | 0.07 | 0.43 | 0.13 | 0.41 |
| | Automated: | 0.40 | 0.07 | 0.40 | 0.09 | 0.00 | 0.43 | 0.07 | 0.43 | 0.13 | 0.00 |
| 12 | **Content Density** | 1.07 | 0.27 | 0.90 | 0.30 | 2.56** | 1.03 | 0.30 | 0.88 | 0.32 | 2.08* |
| | Automated: | 1.04 | 0.27 | 0.87 | 0.30 | 2.56** | 1.01 | 0.29 | 0.84 | 0.32 | 2.39* |
| **Speech Measures** | | | | | | | | | | | |
| 13 | **Pauses per retell** | 8.86 | 5.73 | 9.65 | 4.65 | -0.65 | 7.46 | 4.62 | 6.84 | 4.78 | 0.57 |
| | Automated: | 9.30 | 5.64 | 10.05 | 5.03 | -0.60 | 7.92 | 5.18 | 7.08 | 5.29 | 0.69 |
| 14 | **Total Pause Time** | 21.46 | 21.81 | 18.69 | 12.51 | 0.67 | 11.91 | 10.39 | 11.67 | 11.78 | 0.09 |
| | Automated: | 18.66 | 17.36 | 18.29 | 12.24 | 0.11 | 11.56 | 9.71 | 11.60 | 11.78 | -0.02 |
| 15 | **Mean Pause Duration** | 2.04 | 1.38 | 1.86 | 0.78 | 0.69 | 1.41 | 0.88 | 1.41 | 0.91 | 0.00 |
| | Automated: | 1.80 | 0.94 | 1.76 | 0.73 | 0.20 | 1.32 | 0.80 | 1.35 | 0.78 | -0.16 |
| 16 | **Standardized Pause Rate** | 10.72 | 10.03 | 7.21 | 3.70 | 2.00* | 13.62 | 12.94 | 9.66 | 6.97 | 1.64 |
| | Automated: | 10.93 | 9.49 | 7.50 | 5.46 | 1.91* | 12.08 | 10.18 | 9.76 | 7.40 | 1.12 |
| 17 | **Total Phonation Time** | 24.20 | 9.49 | 20.55 | 8.43 | 1.75* | 24.04 | 9.35 | 19.05 | 10.68 | 2.14* |
| | Automated: | 23.86 | 11.64 | 20.23 | 8.68 | 1.52 | 22.37 | 8.49 | 18.49 | 10.64 | 1.73* |
| 18 | **Phonation Rate** | 0.62 | 0.21 | 0.56 | 0.15 | 1.41 | 0.72 | 0.18 | 0.65 | 0.23 | 1.46 |
| | Automated: | 0.63 | 0.18 | 0.55 | 0.15 | 2.08* | 0.71 | 0.17 | 0.64 | 0.23 | 1.49 |
| 19 | **Transformed Phonation Rate** | 0.93 | 0.25 | 0.85 | 0.16 | 1.64 | 1.04 | 0.23 | 0.94 | 0.31 | 1.58 |
| | Automated: | 0.93 | 0.20 | 0.85 | 0.17 | 1.85* | 1.03 | 0.23 | 0.94 | 0.32 | 1.39 |
| 20 | **Standardized Phonation Time** | 2.98 | 0.61 | 3.01 | 0.58 | -0.22 | 3.06 | 0.67 | 3.02 | 0.88 | 0.22 |
| | Automated: | 3.18 | 0.69 | 3.19 | 0.70 | -0.06 | 3.22 | 0.68 | 3.17 | 0.96 | 0.26 |
| 21 | **Verbal Rate** | 1.88 | 0.79 | 1.69 | 0.57 | 1.19 | 2.21 | 0.80 | 2.07 | 0.82 | 0.74 |
| | Automated: | 2.03 | 0.79 | 1.76 | 0.61 | 1.65 | 2.32 | 0.87 | 2.13 | 0.83 | 0.96 |

more important shorter pauses. The relative contribution of short and long pauses to these measures, therefore, will be the topic of future work.

## E. Classification Trials

We used SVMlight [40] to train support vector machine classifiers using a second-order polynomial kernel and default parameterizations for all trials reported here. Feature values were normalized to fall between 0 and 1 in each training set, by finding the maximum and minimum values for each feature in the training set; the same normalization function was then applied to the left-out pair.[10] The most common method for evaluating classifier quality is the *Receiver Operating Characteristics* (ROC) [41], which plots the sensitivity (true positive rate) of the classifier versus 1−specificity (false positive rate) of the classifier as the classification threshold moves from the most positive threshold (everything true) to the most negative threshold (everything false). A random classifier would have a straight line connecting (0,0) to (1,1). The quality of the classifier is measured as the area under the ROC curve (AUC), also known as the *Wilcoxon-Mann-Whitney* statistic [42], which is an estimate of the probability that a randomly chosen positive example would be ranked higher than a negative example by the classifier. Let $s(e)$ be the score returned by a classifier for a test example $e$. Note that the boolean value $[s(x) > s(y)]$ is 1 if $s(x) > s(y)$ and 0 otherwise. Then the AUC over a set of positive examples $P$ and negative examples $N$ is defined as

$$\mathrm{AUC}(s, P, N) = \frac{1}{|P||N|} \sum_{p \in P} \sum_{n \in N} [s(p) > s(n)]. \qquad (3)$$

Confidence intervals can be defined for this statistic, allowing for meaningful comparison between classifiers. For the current results, we follow [42] and calculate the variance of the AUC, which we denote $A$ below for compactness, using the following formula:

$$\sigma_A^2 = \frac{A(1-A) + (|P|-1)\left(\frac{A}{2-A} - A^2\right) + (|N|-1)\left(\frac{2A^2}{1+A} - A^2\right)}{|P||N|} \qquad (4)$$

We follow a leave-pair-out cross-validation method for calculating the AUC, which is a generalization of the leave-one-out method that is used for cost functions defined over pairs, such as the AUC [43], [44]. Leave-pair-out cross-validation has been shown to provide an unbiased estimate of the error [43]. We use a brute force method to calculate the AUC in this fashion, training a separate classifier for each positive/negative pair, of which there are $(37)^2 = 1369$ in the current case.

[10] Note that test-set feature values transformed with such a function can be negative or greater than 1, hence not strictly constraining the values to $[0, 1]$.

TABLE V
LEAVE-PAIR-OUT CLASSIFICATION RESULTS. SEE TABLES I AND IV FOR TEST AND FEATURE ID NUMBER REFERENCE

| Feature Set | AUC | $\sigma$ |
|---|---|---|
| Automated language features 1-12 + automated speech features 13-21 | 0.627 | 0.649 |
| Automated language features 4, 7-9, 12 + automated speech features 16-19 | 0.703 | 0.061 |
| Automated language features 4, 7-9, 12 | 0.732 | 0.058 |
| Test scores 4, 5 | 0.749 | 0.057 |
| Test scores 4, 5 + automated language features 4, 7-9, 12 | 0.775 | 0.055 |
| Test scores 1-9 | 0.815 | 0.050 |
| Test scores 1-9 + automated language features 4, 7-9, 12 | 0.854 | 0.045 |
| Test scores 1-9 + automated language features 4, 7-9, 12 + automated speech features 16-19 | 0.861 | 0.044 |

In Table V, we present the AUC and standard deviation for classifiers using different feature sets. To be precise about which features were used in which trials, Table V shows feature and test ID numbers, which refer back to features and tests shown in Tables I and IV.

The first three rows of Table V show results using just speech and language derived features, with no test scores. If we use all the automated speech and language derived features from Table IV, the classifier is seriously over-parameterized (row 1); hence, we selected just those features that showed significant group differences in Table IV. The language features thus selected are: words per clause, Yngve per word, dependency length per word, POS cross-entropy (from LM transcripts), and content density, all on Wechsler LM I and II transcripts. The speech features selected were standardized pause rate, total phonation time, phonation rate, and transformed phonation rate, all on both LM I and II retellings. Restricting the classifier to just these features yields a significant improvement (row 2), and in fact restricting to just the significant language features was even better (row 3). Note, however, that this performance is not better than just using the Wechsler Logical Memory test scores (immediate and delayed) by themselves (row 4). Combining the features from rows 3 and 4 yield small improvements in classification (row 5), but not as much as combining scores from all of the tests reported in Table I (row 6). However, combining all test scores with automatically derived language features yields an additional 4% absolute improvement in AUC, and automatically derived speech features provide an additional 0.7%, for a total 4.5% improvement in AUC beyond what was achieved with test scores alone.

## V. DISCUSSION AND FUTURE DIRECTIONS

The results presented in the last section demonstrate that NLP and speech processing techniques applied to clinically elicited spoken language samples can be used to automatically derive measures that may be useful for discriminating between healthy and MCI subjects. In addition, we see that different measures show different patterns when applied to these language samples, with some measures (Yngve, dependency length) showing differences on LM II but not LM I; and others (e.g., POS-tag cross entropy) showing complementary patterns. Given that LM I is an immediate recall test, hence presumably including more verbatim recall among healthy subjects than in the delayed LM II test, such complementary patterns are perhaps not surprising.

The differences in the significance of the Yngve score and dependency length between LM I and LM II seem to be caused by an increase in the average of those scores for unimpaired subjects and a decrease of those scores for impaired subjects across the two retellings. This follows the trend in number of words per retelling. We do not have a definitive explanation for this observed pattern, but it is an interesting signature of MCI.

Three markers that Singh *et al.* found to be significant were not significant in our study, namely Mean Duration of Pauses, Standardized Phonation Time, and Verbal Rate. These markers did not yield a significant difference in either manual or automated annotation. There were two markers that Singh *et al.* did not find significance for, but which they still thought contributed to discrimination between the two groups when combined with other markers. These two markers, Standardized Pause Rate and Phonation Rate, were significant by themselves in our study, particularly for the case of automated annotation. While language features achieved more statistically significant group differences, both feature sets contributed to the improvements in classifier performance presented in Section IV-E.

The differences in the significance of speech-based markers between the Singh *et al.* paper and this paper may be due to the large number of differences between the two studies, including the number of subjects, inclusion criteria for the two subject groups, test material (open-ended questions in the previous study and story retellings in our study), and measurement tools. Perhaps most notable is the difference in mean age of subjects: 61 for healthy and 68 for impaired in their study; 89 for healthy and 90 for impaired in this study. Also, their impaired subjects had been diagnosed as suffering from dementia. Our subjects were in a much earlier stage of cognitive decline characteristic of the MCI syndrome. The markers that were considered informative in both studies, despite these differences, are therefore of particular interest.

In summary, we have demonstrated an important clinical use for NLP and speech-based techniques, where automatic syntactic annotation provides sufficiently accurate parse trees for use in calculation of linguistic complexity measures, and forced alignment provides sufficiently accurate classification of speech and pause regions for use in speech-based measures. Different linguistic complexity measures appear to be measuring complementary characteristics of the retellings, yielding statistically significant differences from both immediate and delayed retellings. Classifiers making use of these features yielded significantly better ROC curve performance than those relying just on summary test scores.

There are quite a number of questions that we will continue to pursue. Most importantly, we will continue to examine this data, to try to determine what characteristics of the spoken language are leading to the unexpected patterns in the results. Eventually, longitudinal tracking of subjects may be the best application of such measures on clinically elicited spoken language samples.

## Acknowledgment

The authors would like to thank to J. van Santen, T. Zitzelberger, J. Payne-Murphy, and R. Guariglia for help with the project.

## References

[1] B. Roark, J. Hosom, M. Mitchell, and J. Kaye, "Automatically derived spoken language markers for detecting mild cognitive impairment," in *Proc. 2nd Int. Conf. Technol. Aging (ICTA)*, 2007.

[2] B. Roark, M. Mitchell, and K. Hollingshead, "Syntactic complexity measures for detecting mild cognitive impairment," in *Proc. ACL 2007 Workshop Biomed. Natural Lang. Process. (BioNLP)*, 2007, pp. 1–8.

[3] K. Sagae, A. Lavie, and B. MacWhinney, "Automatic measurement of syntactic development in child language," in *Proc. 43rd Annu. Meeting ACL*, 2005, pp. 197–204.

[4] K. Gabani, M. Sherman, T. Solorio, Y. Liu, L. Bedore, and E. P. Na, "A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children," in *Proc. Human Lang. Technol. Conf. Meeting North Amer. Chap. ACL (HLT-NAACL)*, 2009, pp. 46–55.

[5] K. Ritchie and J. Touchon, "Mild cognitive impairment: Conceptual basis and current nosological status," *Lancet*, vol. 355, pp. 225–228, 2000.

[6] R. Petersen, "Mild cognitive impairment: Aging to Alzheimer's disease," in *Clinical Features*, R. Petersen, Ed. New York: Oxford Univ. Press, 2003, pp. 15–40.

[7] R. Petersen, G. Smith, S. Waring, R. Ivnik, E. Tangalos, and E. Kokmen, "Mild cognitive impairment: Clinical characterization and outcome," *Archives of Neurol.*, vol. 56, pp. 303–308, 1999.

[8] S. Larrieu, L. Letenneur, J. Orgogozo, C. Fabrigoule, H. Amieva, and N. LeCarret, "Incidence and outcome of mild cognitive impairment in population-based prospective cohort," *Neurology*, vol. 59, pp. 1594–1599, 2002.

[9] D. Howieson, R. Camicioli, J. Quinn, L. Silbert, B. Care, M. Moore, A. Dame, G. Sexton, and J. Kaye, "Natural history of cognitive decline in the old old," *Neurology*, vol. 60, pp. 1489–1494, 2003.

[10] C. DeCarli, "Mild cognitive impairment: Prevalence, prognosis, aetiology and treatment," *Lancet Neurol.*, vol. 2, pp. 15–21, 2003.

[11] J. Morris, "The clinical dementia rating (CDR): Current version and scoring rules," *Neurology*, vol. 43, pp. 2412–2414, 1993.

[12] L. Boise, M. Neal, and J. Kaye, "Dementia assessment in primary care: Results from a study in three managed care systems," *J. Gerontol.*, vol. 59, no. 6, p. M621-6, 2004.

[13] D. Snowdon, S. Kemper, J. Mortimer, L. Greiner, D. Wekstein, and W. Markesbery, "Linguistic ability in early life and cognitive function and Alzheimer's disease in late life," *J. Amer. Med. Assoc.*, vol. 275, no. 7, pp. 528–532, 1996.

[14] S. Kemper, E. LaBarge, F. Ferraro, H. Cheung, H. Cheung, and M. Storandt, "On the preservation of syntax in Alzheimer's disease," *Archives of Neurol.*, vol. 50, pp. 81–86, 1993.

[15] K. Lyons, S. Kemper, E. LaBarge, F. Ferraro, D. Balota, and M. Storandt, "Oral language and Alzheimer's disease: A reduction in syntactic complexity," *Aging and Cogn.*, vol. 1, no. 4, pp. 271–281, 1994.

[16] S. Singh, R. Bucks, and J. Cuerden, "Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech," *Aphasiology*, vol. 15, no. 6, pp. 571–584, 2001.

[17] J. Morris, C. Ernesto, K. Schafer, M. Coats, S. Leon, M. Sano, L. Thal, and P. Woodbury, "Clinical dementia rating training and reliability in multicenter studies: The Alzheimer's disease cooperative study experience," *Neurology*, vol. 48, no. 6, pp. 1508–1510, 1997.

[18] M. Folstein, S. Folstein, and P. McHugh, "'Mini-mental state'—A practical method for grading the cognitive state of patients for the clinician," *J. Psychiatric Res.*, vol. 12, pp. 189–198, 1975.

[19] K. Welsh, N. Butters, R. Mohs, D. Beekly, S. Edland, and G. Fillenbaum, "The consortium to establish a registry for Alzheimer's disease (CERAD part V. A normative study of the neuropsychological battery," *Neurology*, vol. 44, no. 4, pp. 609–614, 1994.

[20] D. Wechsler, *Wechsler Memory Scale—Third Edition Manual*. San Antonio, TX: The Psychological Corp., 1997.

[21] D. Wechsler, *Wechsler Adult Intelligence Scale-Revised*. New York: The Psychological Corp., 1981.

[22] S. Strassel, *Simple Metadata Annotation Specification V6.2*. Philadelphia, PA: Linguistic Data Consortium, 2004.

[23] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: A telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, vol. I, pp. 517–520.

[24] M. Marcus, B. Santorini, and M. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Comput. Linguist.*, vol. 19, no. 2, pp. 313–330, 1993.

[25] V. Yngve, "A model and an hypothesis for language structure," *Proc. Amer. Philosophical Soc.*, vol. 104, pp. 444–466, 1960.

[26] J. Miller and R. Chapman, "The relation between age and mean length of utterance in morphemes," *J. Speech Hearing Res.*, vol. 24, pp. 154–161, 1981.

[27] H. Scarborough, "Index of productive syntax," *Appl. Psycholinguist.*, vol. 11, pp. 1–22, 1990.

[28] S. Rosenberg and L. Abbeduto, "Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults," *Appl. Psycholinguist.*, vol. 8, pp. 19–32, 1987.

[29] L. Frazier, "Syntactic complexity," in *Natural Language Parsing*, D. Dowty, L. Karttunen, and A. Zwicky, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1985, pp. 129–189.

[30] D. Lin, "On the structural complexity of natural language sentences," in *Proc. COLING'96*, 1996, pp. 729–733.

[31] E. Gibson, "Linguistic complexity: Locality of syntactic dependencies," *Cognition*, vol. 68, no. 1, pp. 1–76, 1998.

[32] D. Magerman, "Statistical decision-tree models for parsing," in *Proc. 33rd Annu. Meeting ACL*, 1995, pp. 276–283.

[33] S. Rosenkrantz and P. Lewis, II, "Deterministic left corner parsing," in *Proc. IEEE Conf. Record 11th Annu. Symp. Switching Automata*, 1970, pp. 139–152.

[34] C. Brown, S. Kemper, R. Herman, and M. Covington, "Automatic measurement of propositional density from part-of-speech tagging," *Behavior Res. Methods*, vol. 40, pp. 540–545, 2008.

[35] E. Charniak, "A maximum-entropy-inspired parser," in *Proc. 1st Conf. North Amer. Chap. ACL*, 2000, pp. 132–139.

[36] E. Charniak and M. Johnson, "Edit detection and parsing for transcribed speech," in *Proc. 2nd Conf. North Amer. Chap. ACL*, 2001, pp. 118–126.

[37] E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. P. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski, "A procedure for quantitatively comparing the syntactic coverage of English grammars," in *Proc. DARPA Speech Natural Lang. Workshop*, 1991, pp. 306–311.

[38] M. Bacchiani, M. Riley, B. Roark, and R. Sproat, "MAP adaptation of stochastic grammars," *Comput. Speech Lang.*, vol. 20, no. 1, pp. 41–68, 2006.

[39] J.-P. Hosom, "Speaker-independent phoneme alignment using transitiondependent states," *Speech Commun.*, vol. 51, no. 4, pp. 352–368, 2009.

[40] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.

[41] J. Egan, *Signal Detection Theory and ROC Analysis*. New York: Academic, 1975.

[42] ***[AUTHOR: Please provide page range]***J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, 1982.

[43] C. Cortes, M. Mohri, and A. Rastogi, "An alternative ranking problem for search engines," in *Proc. WEA07, LNCS 4525*, 2007, pp. 1–21, Springer-Verlag.

[44] T. Pahikkala, A. Airola, J. Boberg, and T. Salakoski, "Exact and efficient leave-pair-out cross-validation for ranking RLS," in *Proc. AKRR08*, 2008, pp. 1–8.

**Brian Roark** received the B.A. degree from the Departments of Mathematics and Philosophy, University of California, Berkeley, in 1989, the M.S. degree in information science from the Claremont Graduate School, Claremont, CA, in 1997, the M.S. degree in applied mathematics from Brown University, Providence, RI, in 2001, and the Ph.D. degree from the Department of Cognitive and Linguistic Sciences, Brown University, in 2001.

From 2001 to 2004, he was a Senior Technical Staff Member in the Speech Algorithms Department, AT&T Labs—Research. He is currently an Associate Professor in the Center for Spoken Language Understanding and Department of Biomedical Engineering at Oregon Health and Science University, Portland.

**Margaret Mitchell** received the B.A. degree from the Linguistics Department, Reed College, Portland, OR, in 2005 and the M.A. degree in computational linguistics from the University of Washington, Seattle, in 2008. She is currently pursuing the Ph.D. degree in computing science at the University of Aberdeen, Aberdeen, U.K.

**John-Paul Hosom** received the B.S. degree from the Computer and Information Science Department, University of Massachusetts, Amherst, in 1987 and the Ph.D. degree from the Computer Science and Engineering Department, Oregon Graduate Institute (now part of Oregon Health and Science University), Portland, in 2000.

He is currently an Assistant Professor in the Biomedical Engineering Department, Oregon Health and Science University.

**Kristy Hollingshead** received the B.A. degree from the Department of English, University of Colorado, ***[AUTHOR: Location?]*** in 2000 and the M.S. degree in 2003 and Ph.D. in 2010 from the Department of Computer Science and Engineering, Oregon Health and Science University, Portland, in 2003 and 2010, respectively.

She is currently a Postdoc at the Institute for Advanced Computer Studies, University of Maryland, College Park.

**Jeffrey Kaye** received the M.D. degree from New York Medical College, New York.

From 1981 to 1984, he trained in Neurology at Boston University, Boston, MA, subsequently completing a fellowship in Movement Disorders and Neuropharmacology. He was a Medical Staff Fellow in brain aging at the National Institute on Aging, National Institutes of Health, Bethesda, MD, from 1984 to 1988. He is currently a Professor of Neurology and Biomedical Engineering, the Director of Oregon Center for Aging and Technology, and the Director of the Layton Aging and Alzheimers Disease Center, Oregon Health and Science University, Portland.