Syntactic complexity measures for detecting Mild Cognitive Impairment

Brian Roark, Margaret Mitchell and **Kristy Hollingshead**
Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University
Beaverton, Oregon, 97006 USA
{roark,meg.mitchell,hollingk}@cslu.ogi.edu

Abstract

We consider the diagnostic utility of various syntactic complexity measures when extracted from spoken language samples of healthy and cognitively impaired subjects. We examine measures calculated from manually built parse trees, as well as the same measures calculated from automatic parses. We show statistically significant differences between clinical subject groups for a number of syntactic complexity measures, and these differences are preserved with automatic parsing. Different measures show different patterns for our data set, indicating that using multiple, complementary measures is important for such an application.

1 Introduction

Natural language processing (NLP) techniques are often applied to electronic health records and other clinical datasets. Another potential clinical use of NLP is for processing patient language samples, which can be used to assess language development (Sagae et al., 2005) or the impact of neurodegenerative impairments on speech and language (Roark et al., 2007). In this paper, we present methods for automatically measuring syntactic 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.

Mild Cognitive Impairment (MCI), and in particular amnestic MCI, the earliest clinically defined stage of Alzheimer's-related dementia, often goes undiagnosed due to the inadequacy of common screening tests such as the Mini-Mental State Examination (MMSE) for reliably detecting relatively subtle impairments. 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 correctly recalled) ignore potentially useful information in the characteristics of the spoken language itself.

Narrative retellings provide a natural, conversational speech sample that can be analyzed for many of the characteristics of speech and language that have been shown to discriminate between healthy and impaired subjects, including syntactic complexity (Kemper et al., 1993; Lyons et al., 1994) and mean pause duration (Singh et al., 2001). These measures go beyond simply measuring fidelity to the narrative, thus providing key additional dimensions for improved diagnosis of impairment. Recent work (Roark et al., 2007) has shown significant differences between healthy and MCI groups for both pause related and syntactic complexity measures derived from transcripts and audio of narrative recall tests. In this paper, we look more closely at syntactic complexity measures.

There are two key considerations when choosing how to measure syntactic complexity of spoken language samples for the purpose of psychometric evaluation. First and most importantly, the syntactic complexity measures will be used for discrimination between groups, hence high discriminative utility is desired. It has been demonstrated in past studies (Cheung and Kemper, 1992) that many competing measures are in fact very highly correlated, so it may be the case that many measures are equally discriminative. For this reason, previous results (Roark et al., 2007) have focused on a single syntactic complexity metric, that of Yngve (1960).

A second key consideration, however, is the fidelity of the measure when derived from transcripts via automatic parsing. Different syntactic complexity measures rely on varying levels of detail from

BioNLP 2007: Biological, translational, and clinical language processing, pages 1–8,
Prague, June 2007.

the parse tree. Some syntactic complexity measures, such as that of Yngve (1960), make use of unlabeled tree structures to derive their scores; others, such as that of Frazier (1985), rely on labels within the tree, in addition to the tree structure, to provide the scores. Given these different uses of detail, some measures may be less reliable with automation, hence dis-preferred in the context of automated evaluation. Ideally, simple, easy-to-automate measures with high discriminative utility are preferred.

In the current paper, we demonstrate that various syntactic complexity measures capture complementary systematic differences between subject groups, suggesting that the best approach to discriminating between healthy and impaired subjects is to collect various measures, as a way of capturing language "signatures" of the impairment.

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

2 Syntactic annotation

For manual syntactic annotation of collected data (see Section 4), we followed the syntactic annotation conventions of the well-known Penn Treebank (Marcus et al., 1993). 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, the 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 syntactic complexity measures.

There are characteristics of Penn Treebank annotation that can impact syntactic 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 (1960). Second, in spoken language annotations, a *reparandum*[1] is denoted with a special non-terminal category EDITED. For this paper, we remove from the tree these non-terminals, and the structures underneath them, prior to evaluating syntactic complexity.

3 Syntactic complexity

There is no single agreed-upon measurement of syntactic complexity. 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 (Miller and Chapman, 1981); the Index of Productive Syntax (Scarborough, 1990), a multi-point scale which has recently been automated for child-language transcript analysis (Sagae et al., 2005); and Developmental Level (Rosenberg and Abbeduto, 1987), a 7-point scale of complexity based on the presence of specific grammatical constructions. Other approaches have relied upon the right-branching nature of English syntactic trees (Yngve, 1960; Frazier, 1985), under the assumption that deviations from that correspond to more complexity in the language. Finally, there are approaches focused on the memory demands imposed by "distance" between dependent words (Lin, 1996; Gibson, 1998).

3.1 Yngve scoring

The scoring approach taken in Yngve (1960) is related to the size of a "first in/last out" stack at each word in a top-down, left-to-right parse derivation. Consider the tree in Figure 1. If we knew exactly which productions to use, the parse would begin with an S category on the stack and advance as follows: pop the S and push VP and NP onto the stack; pop NP and push PRP onto the stack; pop PRP from the stack; pop VP from the stack and push NP and VBD onto the stack; and so on. At the point when the word 'she' is encountered, only VP remains on the stack of the parser. When 'was'

[1]A reparandum is a sequence of words that are aborted by the speaker, then *repaired* within the same utterance.

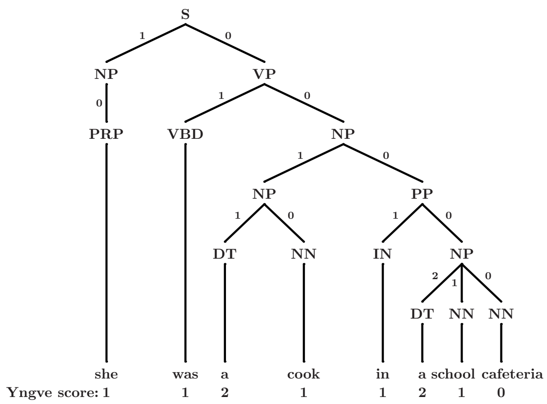

Figure 1: Parse tree with branch scores for Yngve scoring.

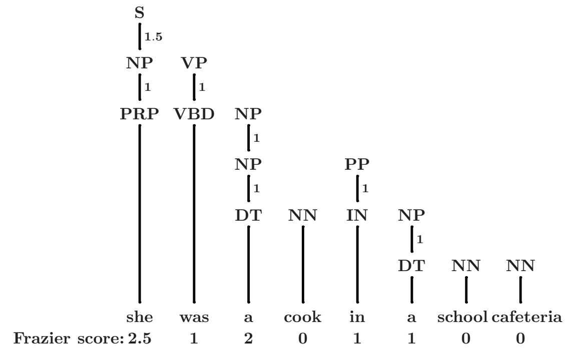

Figure 2: Parse tree fragments with scores for Frazier scoring.

is reached, just NP is on the stack. Thus, the Yngve score for these two words is 1. When the next word 'a' is reached, however, there are two categories on the stack: PP and NN, so this word receives an Yngve score of 2. Stack size has been related by some (Resnik, 1992) to working memory demands, although it most directly measures deviation from right-branching trees.

To calculate the size of the stack at each word, we can use the following simple algorithm. At each node in the tree, label the branches from that node to each of its children, beginning with 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 Figure 1 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.

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 Figure 1, the maximum is 2, the total is 9 and the mean over 8 words is $1\frac{1}{8}$.

3.2 Frazier scoring

Frazier (1985) proposed an approach to scoring syntactic complexity that traces a path from a word up the tree until reaching either the root of the tree or the lowest node which is not the leftmost child of its parent.[2] For example, Figure 2 shows the tree from Figure 1 broken into distinct paths for each word in the string. The first word has a path up to the root, 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,[3] 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 syntactic complexity.

As with the Yngve score, we can calculate the total and the mean of these word scores. In contrast to the maximum score calculated for the Yngve measure, Frazier proposed summing the word scores for each 3-word sequence in the sentence, then taking the maximum of these sums as a measure of highly-localized concentrations of grammatical constituents. For the example in Figure 2, the maximum is 2.5, the maximum 3-word sum is 5.5, and the total is 7.5, yielding a mean of $\frac{15}{16}$.

3.3 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 at-

[2] 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. Empty NPs are annotated in our manual parse trees but not in the automatic parses, which may result in a small disagreement in the Frazier scores for manual and automatic trees.

[3] 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.

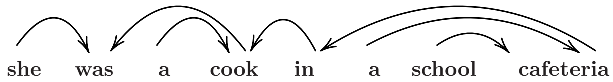

Figure 3: Dependency graph for the example string.

tributed to Magerman (1995), to percolate lexical heads up the tree. Figure 3 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, 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 (1996) 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 (1996) and Gibson (1998) 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 Figure 3. For the current paper, we count the distance between words for each dependency link. For Figure 3, there are 7 dependency links, a distance total of 11, and a mean of $1\frac{4}{7}$.

3.4 Developmental level (D-Level)

D-Level defines eight levels of sentence complexity, from 0-7, based on the development of complex sentences in normal-development children. Each level is defined by the presence of specific grammatical constructions (Rosenberg and Abbeduto, 1987); we follow Cheung and Kemper (1992) in assigning scores equivalent to the defined level of complexity. A score of zero corresponds to simple, single-clause sentences; embedded infinitival clauses get a score of 1 (*She needs to pay the rent*); conjoined clauses (*She worked all day and worried all night*), compound subjects (*The woman and her four children had not eaten for two days*), and wh-predicate complements score 2. Object noun phrase relative clauses or complements score 3 (*The police caught the man who robbed the woman*), whereas the same constructs in subject noun phrases score 5 (*The woman who worked in the cafeteria was robbed*). Gerundive complements and comparatives (*They were hungrier than her*) receive a score of 4; subordinating conjunctions (*if, before, as soon as*) score 6. Finally, a score of 7 is used as a catch-all category for sentences containing more than one of any of these grammatical constructions.

3.5 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.[4]

Given a bi-gram model over POS-tags, we can calculate the probability of the sequence as a whole. Let τ_i be the POS-tag of word w_i in a sequence of words $w_1 \ldots w_n$, and assume that τ_0 is a special start symbol, and that τ_{n+1} is a special stop symbol. Then the probability of the POS-tag sequence is

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

The cross entropy is then calculated as

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

With this formulation, this basically boils down to the mean negative log probability of each tag given the previous tag.

4 Data

4.1 Subjects

We collected audio recordings of 55 neuropsychological examinations administered at the Layton Aging & Alzheimer's Disease Center, an NIA-funded Alzheimer's center for research at OHSU. For this study, we partitioned subjects into two groups: those who were assigned a Clinical Dementia Rating (CDR) of 0 (healthy) and those who were assigned a CDR of 0.5 (Mild Cognitive Impairment; MCI). The CDR (Morris, 1993) is assigned with access to clinical and cognitive test information, independent of performance on the battery of neuropsychological tests used for this research study, and has been shown to have high expert inter-annotator reliability (Morris et al., 1997).

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

Measure	CDR = 0 (n=29) M	SD	CDR = 0.5 (n=18) M	SD	$t(45)$
Age	88.1	9.0	91.9	4.4	-1.65
Education (Y)	15.0	2.2	14.3	2.8	1.04
MMSE	28.4	1.4	25.9	2.6	4.29^{***}
Word List (A)	20.0	4.0	15.4	3.3	4.06^{***}
Word List (R)	6.8	2.0	3.9	1.7	5.12^{***}
Wechsler LM I	17.2	4.0	10.9	4.2	5.20^{***}
Wechsler LM II	15.8	4.3	9.5	5.4	4.45^{***}
Cat.Fluency (A)	17.2	4.1	13.9	4.2	2.59^{*}
Cat.Fluency (V)	12.8	4.5	9.6	3.6	2.57^{*}
Digits (F)	6.6	1.4	6.1	1.2	1.11
Digits (B)	4.7	1.0	4.7	1.1	-0.04

Table 1: Neuropsychological test results for subjects. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Of the collected recordings, three subjects were recorded twice; for the current study only one recording was used for each subject. Three subjects were assigned a CDR of 1.0 and were excluded from the study; two further subjects were excluded for errors in the recording that resulted in missing audio. Of the remaining 47 subjects, 29 had CDR = 0, and 18 had CDR = 0.5.

4.2 Neuropsychological tests

Table 1 presents means and standard deviations for age, years of education and the manually-calculated scores of a number of standard neuropsychological tests that were administered during the recorded session. These tests include: the Mini Mental State Examination (MMSE); the CERAD Word List Acquisition (A) and Recall (R) tests; the Wechsler Logical Memory (LM) I (immediate) and II (delayed) narrative recall tests; Category Fluency, Animals (A) and Vegetables (V); and Digit Span (WAIS-R) forward (F) and backward (B).

The Wechsler Logical Memory I/II tests are the basis of our study on syntactic complexity measures. The original narrative is a short, 3 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). We transcribed each retelling, and manually annotated syntactic parse trees according to the Penn Treebank annotation guidelines. Algorithms for automatically extracting syntactic complexity markers from parse trees were written to accept either manually annotated trees or trees output from an automatic parser, to demonstrate the plausibility of using automatically generated parse trees.

System	LR	LP	F-measure
Out-of-domain (WSJ)	77.7	80.1	78.9
Out-of-domain (SWBD)	84.0	86.2	85.1
Domain adapted from SWBD	87.9	88.3	88.1

Table 2: Parser accuracy on Wechsler Logical Memory responses using just out-of-domain data (either from the Wall St. Journal (WSJ) or Switchboard (SWBD) treebanks) versus using a domain adapted system.

4.3 Parsing

For automatic parsing, we made use of the well-known Charniak parser (Charniak, 2000). Following best practices (Charniak and Johnson, 2001), we removed sequences covered by EDITED nodes in the tree from the strings prior to parsing. For this paper, EDITED nodes were identified from the manual parse, not automatically. Table 2 shows parsing accuracy of our annotated retellings under three parsing model training conditions: 1) trained on approximately 1 million words of Wall St. Journal (WSJ) text; 2) trained on approximately 1 million words of Switchboard (SWBD) corpus telephone conversations; and 3) using domain adaptation techniques starting from the SWBD Treebank. The SWBD out-of-domain system reaches quite respectable accuracies, and domain adaptation achieves 3 percent absolute improvement over that.

For domain adaptation, we used MAP adaptation techniques (Bacchiani et al., 2006) 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 treebank, and the retellings of the other 46 subjects as in-domain training. 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. (2006) for more information on stochastic grammar adaptation using these techniques.

5 Experimental results

5.1 Correlations

Our first set of experimental results regard correlations between measures. Table 3 shows results for five of our measures over all three treebanks that we have been considering: Penn WSJ Treebank, Penn SWBD Treebank, and the Wechsler LM retellings. The correlations along the diagonal are between the same measure when extracted from manually annotated trees and when extracted from automatic

	Penn WSJ Treebank					Penn SWBD Treebank					Wechsler LM Retellings				
	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
(a) Frazier	0.89					0.96					0.94				
(b) Yngve	-0.31	0.96				-0.72	0.96				-0.69	0.95			
(c) Tree nodes	0.91	-0.16	0.92			0.58	-0.06	0.93			0.93	-0.48	0.85		
(d) Dep len	-0.29	0.75	-0.13	0.93		-0.74	0.97	-0.08	0.96		-0.72	0.96	-0.51	0.96	
(e) Cross Ent	0.17	0.18	0.15	0.19	0.93	-0.55	0.76	0.09	0.76	0.98	-0.13	0.45	0.05	0.41	0.97

Table 3: 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.

parses. 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 markers. Interestingly, the number of nodes in the tree per word tends to correlate well with the Frazier score, while the dependency length tends to correlate well with the Yngve score. Cross entropy correlates with Yngve and dependency length for the SWBD and Wechsler treebanks, but not for the WSJ treebank.

5.2 Manually derived measures

Table 4 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. The first three measures presented in the table are available without syntactic 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. The number of clauses are automatically extracted from the parses by counting the number of S nodes in the tree.[5] 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 LM I though not for LM II.

The other measures are as described in Section 3. Interestingly, Frazier score per word, the number of tree nodes per word, and POS-tag cross entropy all show a significant negative t-value on the LM I retellings, meaning the CDR 0.5 subjects had significantly higher scores than the CDR 0 subjects for these measures on this task. These measures showed no significant difference on the LM II retellings.

The Yngve score per word and the dependency length per word showed no significant difference on LM I retellings but a statistically significant difference on LM II, with the expected outcome of higher scores for the CDR 0 subjects. The D-Level measure showed no significant differences.

5.3 Automatically derived measures

In addition to manual-parse derived measures, Table 4 also presents the same measures when automatic, rather than manual, parses are used. Given the relatively high quality of the automatic parses, most of the means and standard deviations are quite close, and all of the patterns observed in the upper half of Table 4 are preserved, except that the Yngve score per word no longer shows a statistically significant difference for the LM II retelling.

5.4 Left-corner trees

For the tree-based complexity metrics (Frazier and Yngve), we also investigated alternative implementations that make use of the left-corner transformation (Rosenkrantz and Lewis II, 1970) of the tree from which the measures were extracted. This transformation is widely known for removing left-recursion from a context-free grammar, and it changes the tree shape by transforming left-branching structures into right-branching structures, while leaving center-embedded structures center-embedded. This property led Resnik (1992) to propose left-corner processing as a plausible mechanism for human sentence processing, since it is precisely these center-embedded structures, and not the left- or right-branching structures, that are problematic for humans to process.

Table 5 presents results using either manually annotated trees or automatic parses to extract the Yngve and Frazier measures after a left-corner transform has been applied to the tree. The Frazier scores are very similar to those without the left-

[5]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.

Measure	Logical Memory I CDR = 0 M	SD	CDR = 0.5 M	SD	$t(45)$	Logical Memory II CDR = 0 M	SD	CDR = 0.5 M	SD	$t(45)$
Total words in retelling	71.0	26.0	58.1	31.9	1.49	70.6	21.5	58.5	36.7	1.43
Total utterances in retelling	8.86	4.16	7.72	3.28	0.99	8.17	2.77	7.06	4.86	1.01
Words per utterance in retelling	8.57	2.44	7.78	3.67	0.89	9.16	3.06	7.82	4.76	1.18
Manually extracted: Words per clause	6.33	1.39	5.25	1.25	2.68^{*}	6.12	1.20	5.48	3.37	0.93
Frazier score per word	1.19	0.09	1.26	0.11	-2.68^{*}	1.19	0.09	1.13	0.43	0.67
Tree nodes per word	1.96	0.07	2.01	0.10	-2.08^{*}	1.96	0.07	1.80	0.66	1.36
Yngve score per word	1.44	0.23	1.39	0.30	0.61	1.53	0.27	1.26	0.62	2.01^{*}
Dependency length per word	1.54	0.25	1.47	0.27	0.90	1.63	0.30	1.34	0.60	2.19^{*}
POS-tag Cross Entropy	1.83	0.16	1.96	0.26	-2.18^{*}	1.93	0.14	1.86	0.59	0.54
D-Level	1.07	0.75	1.03	1.23	0.14	1.23	0.81	1.68	1.41	-1.42
Auto extracted: Words per clause	6.42	1.53	5.10	1.16	3.13^{**}	6.04	1.25	5.61	3.67	0.59
Frazier score per word	1.16	0.10	1.24	0.10	-2.92^{**}	1.15	0.10	1.09	0.41	0.69
Tree nodes per word	1.96	0.07	2.03	0.10	-2.55^{*}	1.96	0.08	1.79	0.66	1.38
Yngve score per word	1.41	0.23	1.37	0.29	0.54	1.50	0.27	1.28	0.64	1.70
Dependency length per word	1.51	0.25	1.47	0.28	0.54	1.61	0.28	1.35	0.61	2.04^{*}
POS-tag Cross Entropy	1.83	0.17	1.96	0.26	-2.12^{*}	1.92	0.14	1.86	0.58	0.53
D-Level	1.09	0.73	1.11	1.20	-0.08	1.28	0.77	1.61	1.22	-1.15

Table 4: Syntactic complexity measure group differences when measures are derived from either manual or automatic parse trees. $^{**}p < 0.01$; $^{*}p < 0.05$

corner transform, while the Yngve scores are reduced across the board. With the left-corner transformed tree, the automatically derived Yngve measure retains the statistically significant difference shown by the manually derived measure.

6 Discussion and future directions

The results presented in the last section demonstrate that NLP 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 Frazier scores and tree nodes per word giving quite different results than Yngve scores and dependency length. It would thus appear that, for Penn Treebank style annotations at least, these measures are quite complementary.

There are two surprising aspects of these results: the significantly higher means of three measures on LM I samples for MCI subjects, and the fact that one set of measures show significant differences on LM I while another shows differences on LM II. We do not have definitive explanations for these phenomena, but we can speculate about why such results were obtained.

First, there is an important difference between the manner of elicitation for LM I versus LM II. LM I is an immediate recall, so there will likely be, for unimpaired subjects, much higher verbatim recall of the story than in the delayed recall of LM II. For the MCI group, which exhibits memory impairment, there will be little in the way of verbatim recall, and potentially much more in the way of spoken language phenomena such as filled pauses, parentheticals and off-topic utterances. This may account for the higher Frazier score per word for the MCI group on LM I. Such differences will likely be lessened in the delayed recall.

Second, the Frazier and Yngve metrics differ in how they score long, flat phrases, such as typical base NPs. Consider the ternary NP in Figure 1. The first word in that NP ('a') receives an Yngve score of 2, but a Frazier score of only 1 (Figure 2), while the second word in the NP receives an Yngve score of 1 and a Frazier score of 0. For a flat NP with 5 children, that difference would be 4 to 1 for the first child, 3 to 0 for the second child, and so forth. This difference in scoring relatively common syntactic constructions, even those which may not affect human memory load, may account for such different scores achieved with these different measures.

In summary, we have demonstrated an important clinical use for NLP techniques, where automatic syntactic annotation provides sufficiently accurate parse trees for use in automatic extraction of syntactic complexity measures. Different syntactic complexity measures appear to be measuring quite complementary characteristics of the retellings, yielding statistically significant differences from both immediate and delayed retellings.

There are quite a number of questions that we will

Measure		Logical Memory I CDR = 0 M	SD	CDR = 0.5 M	SD	$t(45)$	Logical Memory II CDR = 0 M	SD	CDR = 0.5 M	SD	$t(45)$
Manually extracted:	Left-corner Frazier	1.20	0.10	1.28	0.12	-2.60^{*}	1.20	0.11	1.18	0.45	0.29
	Left-corner Yngve	1.33	0.20	1.25	0.23	1.20	1.37	0.21	1.14	0.52	2.14^{*}
Auto extracted:	Left-corner Frazier	1.16	0.10	1.27	0.13	-3.02^{**}	1.15	0.11	1.10	0.42	0.64
	Left-corner Yngve	1.31	0.19	1.23	0.21	1.33	1.36	0.21	1.13	0.51	2.11^{*}

Table 5: Syntactic complexity measure group differences when measures are derived from left-corner parse trees. $^{**}p < 0.01$; $^{*}p < 0.05$

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. In addition, we will begin to explore composite measures, such as differences in measures between LM I and LM II, which promise to better capture some of the patterns we have observed. Ultimately, we would like to build classifiers making use of a range of measures as features, although in order to demonstrate statistically significant differences between classifiers, we will need much more data than we currently have. Eventually, longitudinal tracking of subjects may be the best application of such measures on clinically elicited spoken language samples.

Acknowledgments

This research was supported in part by NSF Grant #IIS-0447214 and pilot grants from the Oregon Center for Aging & Technology (ORCATECH, NIH #1P30AG024978-01) and the Oregon Partnership for Alzheimer's Research. Also, the third author of this paper was supported under an NSF Graduate Research Fellowship. 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. Thanks to Jeff Kaye, John-Paul Hosom, Jan van Santen, Tracy Zitzelberger, Jessica Payne-Murphy and Robin Guariglia for help with the project.

References

M. Bacchiani, M. Riley, B. Roark, and R. Sproat. 2006. MAP adaptation of stochastic grammars. *Computer Speech and Language*, 20(1):41–68.

E. Charniak and M. Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the 2nd Conference of the North American Chapter of the ACL*.

E. Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the ACL*, pages 132–139.

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

L. Frazier. 1985. Syntactic complexity. In D.R. Dowty, L. Karttunen, and A.M. Zwicky, editors, *Natural Language Parsing*. Cambridge University Press, Cambridge, UK.

E. Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

S. Kemper, E. LaBarge, F.R. Ferraro, H. Cheung, H. Cheung, and M. Storandt. 1993. On the preservation of syntax in Alzheimer's disease. *Archives of Neurology*, 50:81–86.

D. Lin. 1996. On structural complexity. In *Proceedings of COLING-96*.

K. Lyons, S. Kemper, E. LaBarge, F.R. Ferraro, D. Balota, and M. Storandt. 1994. Oral language and Alzheimer's disease: A reduction in syntactic complexity. *Aging and Cognition*, 1(4):271–281.

D.M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 276–283.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

J.F. Miller and R.S. Chapman. 1981. The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research*, 24:154–161.

J. Morris, C. Ernesto, K. Schafer, M. Coats, S. Leon, M. Sano, L. Thal, and P. Woodbury. 1997. Clinical dementia rating training and reliability in multicenter studies: The Alzheimer's disease cooperative study experience. *Neurology*, 48(6):1508–1510.

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

P. Resnik. 1992. Left-corner parsing and psychological plausibility. In *Proceedings of COLING-92*, pages 191–197.

B. Roark, J.P. Hosom, M. Mitchell, and J.A. Kaye. 2007. Automatically derived spoken language markers for detecting mild cognitive impairment. In *Proceedings of the 2nd International Conference on Technology and Aging (ICTA)*.

S. Rosenberg and L. Abbeduto. 1987. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8:19–32.

S.J. Rosenkrantz and P.M. Lewis II. 1970. Deterministic left corner parsing. In *IEEE Conference Record of the 11th Annual Symposium on Switching and Automata*, pages 139–152.

K. Sagae, A. Lavie, and B. MacWhinney. 2005. Automatic measurement of syntactic development in child langugage. In *Proceedings of the 43rd Annual Meeting of the ACL*.

H.S. Scarborough. 1990. Index of productive syntax. *Applied Psycholinguistics*, 11:1–22.

S. Singh, R.S. Bucks, and J.M. Cuerden. 2001. Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech. *Aphasiology*, 15(6):571–584.

V.H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.