Applied Psycholinguistics 13 (1992), 53–76
Printed in the United States of America

Competing complexity metrics and adults' production of complex sentences

HINTAT CHEUNG and SUSAN KEMPER
University of Kansas

ADDRESS FOR CORRESPONDENCE
Susan Kemper, 1082 Robert Dole Human Development Center, Child Language Program,
University of Kansas, Lawrence, KS 66045

ABSTRACT
The adequacy of 11 metrics for measuring linguistic complexity was evaluated by applying each metric to language samples obtained from 30 different adult speakers, aged 60–90 years. The analysis then determined how well each metric indexed age-group differences in complexity. In addition, individual differences in the complexity of adults' language were examined as a function of these complexity metrics using structural equation modeling techniques. In a follow-up study, judges listened to sentences in noise, rated their comprehensibility, and attempted to recall each sentence verbatim. Hierarchical multiple regression was used to evaluate the structural equation model, derived from the language samples, with respect to sentence comprehensibility and recall. While most of the metrics provided an adequate account of age-group and individual differences in complexity, the amount of embedding and the type of embedding proved to predict how easily sentences are understood and how accurately they are recalled.

Psycholinguists have frequently attempted to formulate ways of measuring the complexity of different sentences (Brown, 1973; Crain & Schankweiler, 1988; Fay, 1980; Ford, 1983; Frazier, 1988; Lee, 1974; Scarborough, 1990; Smith, 1988; Watt, 1970). Complexity metrics have been theoretically derived from specific linguistic theories, experimentally devised from models of syntactic processing, and empirically developed from research on language acquisition and sentence processing. Complexity metrics are important research tools since they enable researchers to order experimental stimuli from least to most complex, examine developmental trends in children's mastery of grammatical constructions, make cross-linguistic comparisons as to the relative complexity of grammatical constructions in different languages, and so on.

The most widely known attempt to develop a complexity metric was labeled by Fodor, Bever, and Garrett (1974) as the Derivational Theory of Complexity (DTC). The DTC tried to equate the complexity of sentences with the number of transformations, required in the then-current model of generative transformational grammar, intervening between the sentence's deep and surface structures. Experimental findings, which indicated that not all transformations increase processing complexity (see Fodor et al., 1974, for a review), and theoretical changes in syntactic theory (Bresnan, 1982) led to the abandonment of the DTC.

Nonetheless, a variety of approaches to measuring syntactic complexity have been undertaken since the abandonment of the DTC. The most common approach does not postulate a general complexity metric, but rather contrasts children's or adults' processing of alternative grammatical constructions (see, for example, Clancy, Lee, & Zoh, 1986; Frazier & Fodor, 1978; Shapiro, Zurif, & Grimshaw, 1987; Smith & van Kleeck, 1986). Other researchers have attempted to develop metrics that will generally apply to sentences in order to scale the sentences as to their relative difficulty for production or comprehension. Some, like mean length of utterance (MLU) (Brown, 1973) and mean clauses per utterance (MCU) (Kemper, Kynette, Rash, Sprott, & O'Brien, 1989) measure sentence length and assume that sentence length - whether measured in morphemes, words, or clauses - indexes complexity. Other metrics, like those of Botel and Granowsky (1972), Lee (1974), and Rosenberg and Abbeduto (1987), establish an ordering of grammatical constructions based on developmental patterns or the frequency of occurrence of target constructions in speech. A final class of metrics examine structural aspects of sentences and attempt to quantify the processing demands of various sentence structures (Frazier, 1985; Yngve, 1960).

The present investigation of the utility of different complexity metrics was undertaken as part of a study of age-group and individual differences in adults' language. Kemper (1988) suggested that there is an age-related decline in the complexity of adults' language. Previous research (Kynette & Kemper, 1986) established that elderly adults in their 70s and 80s are less likely than younger adults to produce sentences with multiple embedded clauses. MCU appears to decline with advancing age for written diary entries (Kemper, 1987a), spontaneous statements (Kemper et al., 1989), and oral narratives (Kemper, Rash, Kynette, & Norman, 1990), although there are genre differences in the overall incidence of sentence embedding. Of particular interest is the finding that the age-group decline in sentence embedding is somewhat more precipitous for sentences with left-branching structures, including those with sentence-initial subordinate clauses, *that* clauses and *wh-* clauses as subjects, and relative clauses modifying the sentence subject, than for sentences with right-branching structures, such as those with sentence-final subordinate clauses, verb phrase infinitive complements, or relative clauses modifying the sentence predicate.

Kemper and Rash (1988) and Kemper (1988) linked this asymmetry in the production of left- and right-branching sentences to similar asymmetries in processing left- and right-branching sentences; elderly adults have more difficulty recalling (Kemper, 1987b) and imitating (Kemper, 1986) left-branching sentences. Left-branching sentences are presumed to be more difficult to process (Fodor, Bever, & Garrett, 1974) because they impose more demands on working memory to retain and manipulate grammatical constituents than do right-branching sentences.[1]

As evidence for this linkage between working memory and the production of embedded sentences, Kemper and Rash (1988) computed the Yngve depth (Yngve, 1960) of a sample of adults' sentences. Yngve (1960) assumed that the production of a sentence imposed demands on a limited-capacity working memory in order to retain planned but not yet articulated grammatical constituents. The depth of any word in a sentence represents the number of planned grammatical constituents that have not yet been realized during the left-to-right production of the sentence. In general, sentence embedding, particularly left-branching embedding, increases the Yngve depth of sentences since words within the embedded sentence are at greater depth than words in the main clause.

Kemper and Rash (1988) showed that Yngve depth declines with the age of the speaker. They also found that Yngve depth is correlated with adults' backward digit span (Wechsler, 1958). Kemper et al. (1989) found that adults' backward digit is correlated with MCU and the production of left-branching clauses; adults with larger backward digit spans produce sentences with more embedded clauses, particularly left-branching clauses, and greater Yngve depth. This finding implies that the age-related decline in adults' production of complex sentences, particularly left-branching sentences, is due to age-related declines in the capacity of working memory, as measured by backward digit span.

Frazier (1985) challenged Yngve depth as a valid measure of syntactic complexity and suggested an alternative metric which was explicitly motivated by considerations of the complexity of sentence-processing operations. Frazier's account differs from Yngve depth in two ways: first, sentence embeddings are explicitly acknowledged as sources of complexity and, hence, increase the complexity of a particular sentence; second, the complexity is computed over three-word sequences such that a cluster of many processing decisions contributes more to the complexity of a sentence than a distributed sequence of processing decisions.

Experiment 1 was undertaken in order to compare Yngve depth to Frazier's alternative as well as to other complexity metrics. Following a survey of the literature on language processing and language acquisition, a set of 11 complexity metrics was chosen according to two criteria. First, clear rules or procedures for computing each metric were given in the original source, and second, each metric was, in principle, applicable to a wide range of sentences. This last criterion excluded metrics that apply to limited types of sentences such as relative clauses (Clancy, Lee, & Zoh, 1986) or multiclause sentences with missing complement subjects (Hsu, Cairns, & Fiengo, 1985).

The measures were: MLU in words, traditionally used in the child language literature to measure linguistic development (Miller & Chapman, 1981); MCU, used by Kemper et al. (1989) to measure adults' linguistic development; Developmental Sentence Scoring (DSS), developed by Lee (1974) to assess children's grammatical development; Index of Productive

Syntax (IPSyn), which was derived by Scarborough (1990) to scale children's grammatical development; Developmental Level (DLevel), used by Rosenberg and Abbeduto (1987) to evaluate the grammatical competence of retarded adults; Directional Complexity (DComplexity), based on the Botel and Granowsky (1972) formula (developed as an alternative to readability formulas) to measure the linguistic difficulty of texts; two alternative ways of measuring Yngve depth; and two variants of Frazier's node count. In addition, Propositional Density (PDensity), based on Kintsch and Keenan's (1973) analyses of text difficulty, was also computed in order to assess whether semantic content covaries with grammatical complexity.

Experiment 1 was designed to compare the reliability of these complexity metrics and their utility as models of language change in adulthood. Each metric was evaluated as to its adequacy for describing both age-group and individual differences in linguistic complexity. Experiment 2 then provided converging evidence as to the selection of an adequate complexity metric for predicting the relative comprehensibility and verbatim recall of sentences.

EXPERIMENT 1

In order to compare alternative ways of measuring linguistic complexity, 11 different complexity measures were applied to language samples obtained from 30 different adults. The analysis then determined how well each measure indexed age-group differences in complexity. Finally, individual differences in linguistic complexity were examined as a function of each metric. At issue was which metric(s) would provide the best account of both age-group and individual differences in complexity.

The metrics differ in three regards. First, some of the metrics are sensitive to sentence length: MLU, obviously, provides a measure of sentence length, and one each of the Yngve and Frazier metrics must necessarily increase as sentences increase in length, since these metrics are computed by summing scores assigned to each word in a sentence. To the extent that syntactically complex constructions involve more words and more word types, such as complementizers and subordinating conjunctions, MCU, DLevel, DSS, and DComplexity will also increase with sentence length. Hence, one issue is whether there are age-group and individual differences in complexity when sentence length (or MLU) is held constant.

Second, several of the measures explicitly assume that some grammatical constructions are more complex than others; DSS, DComplexity, DLevel, and both Frazier metrics award more points per sentence to embedded clauses (particularly those producing left-branching structures) and subordinate clauses. For these metrics, multiple levels of sentence embedding and subordination must lead to higher scores. Thus, a second issue is whether age-group and individual differences in complexity due to the occurrence of particular types of embedding and subordination will be found even when MCU (or the amount of embedding and subordination per se) is held constant.

Finally, 10 of the metrics assess differences in grammatical form, whereas

PDensity attempts to measure information load (or semantic content). The third issue is whether age-group and individual differences in complexity will be obained when PDensity, or semantic content, is held constant.

The relationships among amount of embedding, type of embedding, and semantic content as alternative sources of complexity were evaluated by comparing a series of structural equation models of the data.

Method

Language samples. The language samples were taken from the oral narratives analyzed by Kemper et al. (1990). The narratives were collected from adults aged 60–90 years. Ten narratives were selected from each age group with the requirement that each contained at least 50 sentences; short narratives containing less than 50 sentences were excluded. There were 10 narratives from adults aged 60 to 69 years, 10 from adults aged 70 to 79 years, and 10 from adults aged 80 to 90 years. Each narrative was told by a native speaker of English. The vocabulary score of each speaker and each speaker's forward and backward digit span scores from the Wechsler Adult Intelligence Scales (WAIS) (Wechsler, 1958) were available. For this sample, age correlated, $r(28) = -.54$, $p < .01$ (two-tailed), with backward digit span, and $r(28) = -.32$, $p < .05$, with forward digit span. The two digit span measures were correlated, $r(28) = +.91$, $p < .01$. Age was not significantly correlated, $r(28) = +.27$, $p > .05$, with vocabulary, nor was vocabulary significantly correlated with forward or backward digit span, both $r(28) < \pm.12$, $p > .05$. The speakers' educational level (years of formal education completed) was also available. Educational level correlated, $r(28) = +.44$, $p < .01$, with vocabulary, but educational level was not correlated with age or forward or backward digit span, all $r(28) < \pm.12$, $p > .10$.

Analyses. A total of 1,500 sentences was analyzed. Fifty consecutive sentences were selected from each language sample for analysis. Only complete sentences were analyzed. Eleven different complexity measures were then obtained for each sentence:

1. MLU. The number of words per sentence was determined and each speaker's MLU was calculated.

2. MCU. The number of syntactic clauses per sentence was determined by counting each main clause and each embedded or subordinate clause. Each speaker's MCU was calculated.

3. DSS. Eight different categories of grammatical forms were scored: indefinite pronouns, personal pronouns, main verbs, secondary (embedded verbs), conjunctions, negatives, and two types of questions. Within each category, variants were assigned different points to reflect the developmental order of appearance in children's speech. A total score was derived for each sentence by summing the points for each category plus 1 point if the sentence was fully grammatical. The average DSS for each speaker was determined based on the sample of 50 sentences.

4. IPSyn. Unlike the other metrics, which apply to grammatical tokens (individual sentences), IPSyn is based on the analysis of grammatical types. This metric is a summary score of how many of 56 target grammatical types each speaker produces. A maximum of two occurrences of each type is tallied; maximal IPSyn score is, therefore, 112. IPSyn grammatical types include: 11 different types of noun phrases, 16 different types of verb phrases, 10 different types of questions (which rarely occurred in the narratives), and 19 different types of sentence structures. The IPSyn score of each speaker was determined for the 50 sentences.

5. DLevel. Eight developmental levels were used to classify the sentences. The original scale, developed by Rosenberg and Abbeduto (1987) specified seven levels of complex sentences; a zero level was added to this system to classify simple, one-clause sentences (42% of the corpus). The eight levels, therefore, were: (0) simple, one-clause sentences; (1) complex sentences with embedded infinitival complements, (2) complex sentences with *wh*-predicate complements, conjoined clauses, and compound subjects, (3) complex sentences with relative clauses modifying the object noun phrase or with predicate noun phrase complements, (4) complex sentences with gerundive complements or comparative constructions, (5) complex sentences with relative clauses modifying the subject noun phrase, subject noun phrase complements, and subject nominalizations, (6) complex sentences with subordinate clauses, and (7) complex sentences with multiple forms of embedding and subordination. The average DLevel for each speaker was calculated.

6. DComplexity. The rules given by Botel and Granowsky (1972) were applied to each sentence to determine DComplexity. These rules assign 0, 1, 2, or 3 points to various sentence patterns and structures. 0-point structures include subject–verb, subject–verb–object, and subject–verb–infinitive constructions; interrogative sentences; and coordinate clauses joined by *and*. 1-point structures include sentences with both direct and indirect objects; noun modifiers such as adjectives and possessives; adverbials; coordinate clauses joined by *but*, *or*, and so forth; gerunds used as subjects; and infinitive complements to subject–verb–object clauses. 2-point structures include comparatives, subordinate clauses, infinitives used as subjects, and passives. 3-point structures include *wh*- and *that* clauses used as subjects. The average DComplexity of each speaker's utterances was calculated.

7. and 8. Yngve depth. Both the total Yngve depth and the maximum Yngve depth of each sentence were determined according to the procedures given by Yngve (1960). Figures 1 and 2 illustrate the calculation of the Yngve depth measures. Yngve depth was determined by first performing a surface phrase structure analysis of the sentence to construct a syntactic tree with nodes and branches, and then by numbering the branches below each node from right to left, starting with zero. The depth of each word was the sum of all the branches connecting the word to the root or top-most node of the sentence.

In performing this analysis, a surface phrase structure approach (McCaw-

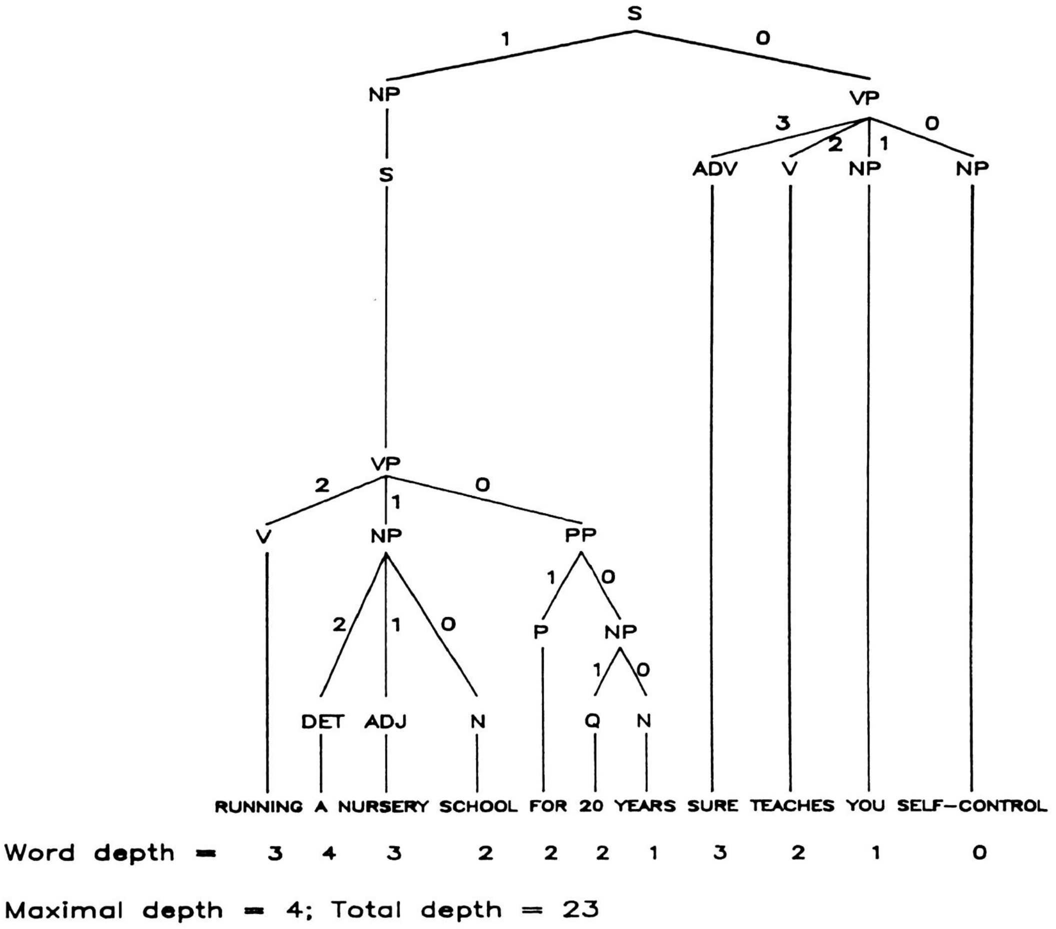
S
1
0
NP
VP
S
3
2
1
0
ADV
V
NP
NP
VP
2
1
0
V
NP
PP
2
1
0
1
0
P
NP
1
0
DET
ADJ
N
Q
N
RUNNING A NURSERY SCHOOL FOR 20 YEARS SURE TEACHES YOU SELF–CONTROL
Word depth = 3 4 3 2 2 2 1 3 2 1 0
Maximal depth = 4; Total depth = 23

Figure 1. Yngve analysis of a left-branching sentence, showing both maximal and total Yngve depths.

ley, 1989) was used to parse the sentences. Note that an inflated Yngve depth measure would be produced if, for example, the bar-X notation (Jackendoff, 1977) was employed. Thus, "the big girl" was analyzed as $[_{NP}$ the $_{DET}$ big $_{ADJ}$ girl $_{N}]$ rather than as $[_{N}$ the $_{SPEC}$ $[_{N}$ big $_{SPEC}$ $[_{N}$ girl$]]]$.

Many of the speakers' sentences began with conjunctions, usually *and*; consequently, Yngve depth would also be inflated by treating these sentence-initial conjunctions as branches originating from the root of the tree structure. Such conjunctions inflate Yngve depth by 1 since three branches (i.e., the conjunction, the subject noun phrase, and the verb phrase) originate from the root of the tree. To avoid inflating Yngve depth, sentence-initial conjunctions were not included in its computation.

Two Yngve depth measures were determined for each sentence: (a) Maximal Yngve depth is the largest number associated with any word in the sentence, and (b) Total Yngve depth is the sum of all depth counts for each word in the sentence. Maximal Yngve depth was, therefore, a "local" measure that was independent of sentence length, whereas Total Yngve

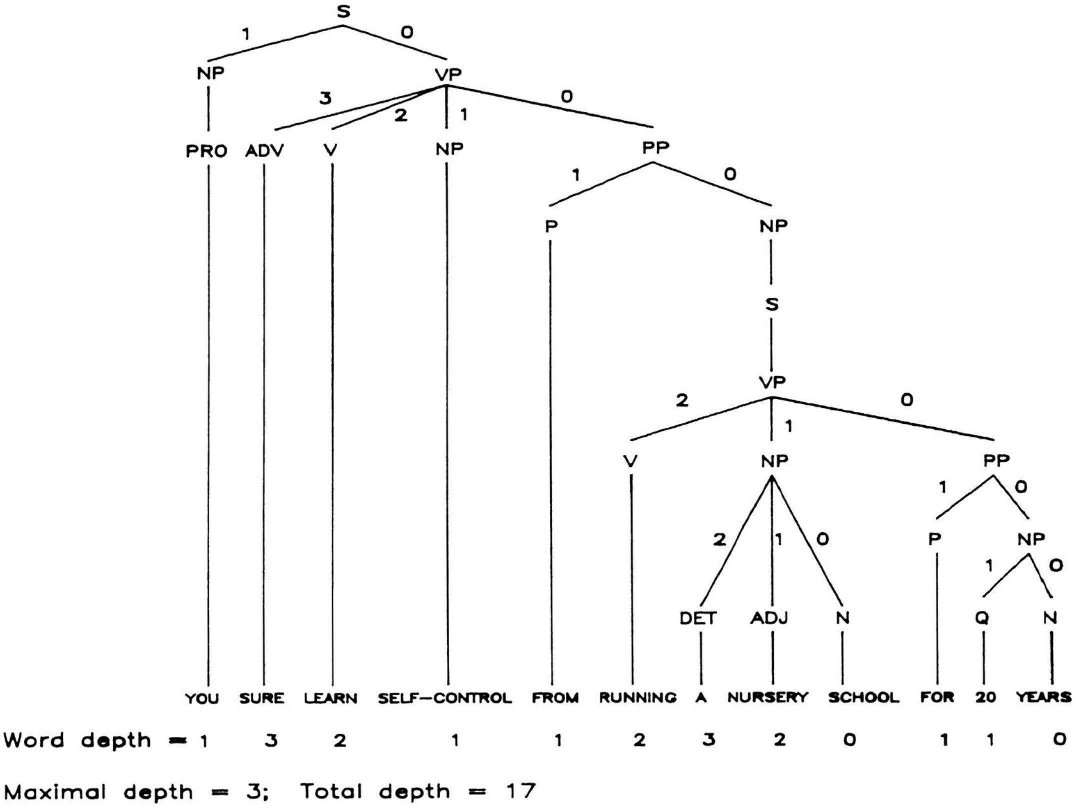

Figure 2. Yngve analysis of a right-branching sentence, showing both maximal and total Yngve depths.

depth was confounded with the number of words in the sentence. The average Maximum Yngve depth and the average Total Yngve depth were computed for each speaker.

9. and 10. Frazier count. Two measures, Local Frazier node count and Total Frazier node count, were derived from the rules given by Frazier (1985). Figures 3 and 4 illustrate the calculation of the Local Frazier and Total Frazier counts. The Frazier counts were based on a surface phrase structure analysis in which all (nonterminal) nodes in the phrase structure of the sentence were assigned a point value of 1 except for sentence nodes and sentence-complement nodes, which were assigned a point value of 1.5. Counts for each word were then determined by summing up the points assigned to all the nodes dominating each word in the sentence.

As implied by the analyses given in Frazier (1985), nodes in the phrase structure of a sentence were counted as if the sentence was being parsed from left to right in a deterministic manner, as in the parser developed by Marcus (1980) and discussed by Berwick and Weinberg (1984). Consequently, nodes were assigned to possessive markers and deleted noun phrases that introduce new syntactic constituents or that are required in order to connect each new word to the preceding structure. For example, a gerund is used as the subject of the main sentence in the left-branching

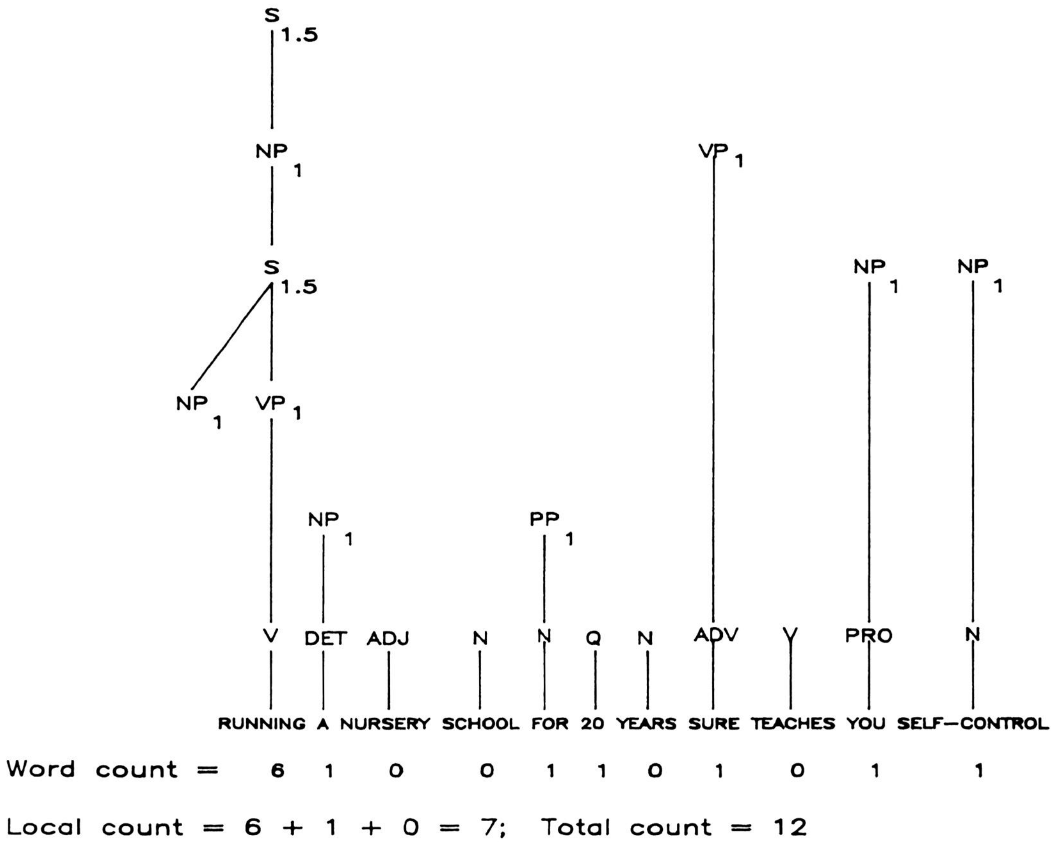
S 1.5
NP 1
S 1.5
NP 1
VP 1
NP 1
PP 1
VP 1
NP 1
NP 1
V
DET
ADJ
N
N
Q
N
ADV
V
PRO
N
RUNNING A NURSERY SCHOOL FOR 20 YEARS SURE TEACHES YOU SELF—CONTROL
Word count = 6 1 0 0 1 1 0 1 0 1 1
Local count = 6 + 1 + 0 = 7; Total count = 12

Figure 3. Frazier analysis of a left-branching sentence, showing both local and total Frazier counts.

version (Figure 3) and as the object of a preposition in the right-branching version (Figure 4); in either version, the subject of the embedded verb has been deleted, and the entire gerund functions as a noun phrase. The gerund is given 1 point as a noun phrase, 1.5 points for the embedded S node, 1 point for the empty NP subject node, and 1 point since it is a verb; in the left-branching version (Figure 3), it also receives 1.5 points for the main S node since the gerund is the subject of the main sentence.

Two variants of the Frazier count were computed. The Local Frazier count was determined by summing the node points for each sequence of three adjacent words and identifying the largest such sum in the sentence. This three-word window is assumed to reflect the capacity of the parser to hold partially analyzed constituents (Marcus, 1980) and thus represents a cluster of many processing decisions. The Total Frazier count was determined by summing all node points for all the words in each sentence. The Local Frazier count, therefore, reflects a concentration of grammatical constituents, whereas the Total Frazier count was confounded with the length of the sentence. Average Local Frazier and Total Frazier counts were obtained for each speaker.

11. PDensity. Using the procedures given by Turner and Greene (1977),

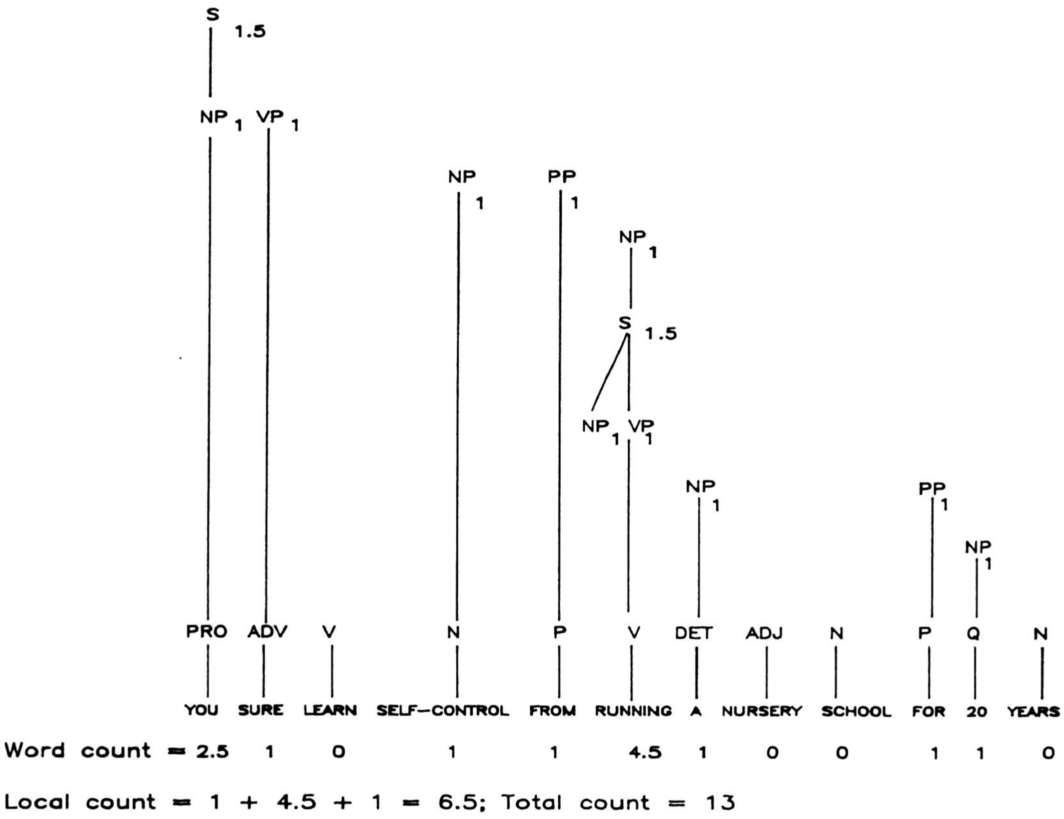
S 1.5
NP 1 VP 1
NP 1
PP 1
NP 1
S 1.5
NP 1 VP 1
NP 1
PP 1
NP 1
PRO ADV V N P V DET ADJ N P Q N
YOU SURE LEARN SELF—CONTROL FROM RUNNING A NURSERY SCHOOL FOR 20 YEARS
Word count = 2.5 1 0 1 1 4.5 1 0 0 1 1 0
Local count = 1 + 4.5 + 1 = 6.5; Total count = 13

Figure 4. Frazier analysis of a right-branching sentence, showing both local and total Frazier counts.

each sentence was decomposed into a set of underlying propositions. Propositions are "idea units" and fall into three classes: predicates expressing actions or states; modifications expressing restrictions or limitations, including qualifications, quantifications, and negations; and connections, including conjunction, disjunction, causality, and contrast. The work of Kintsch and his colleagues (Kintsch & Keenan, 1973; Kintsch & Vipond, 1978) suggested that propositional density (PDensity), or the number of propositions per 100 words of text, is a determinate of reading difficulty. It can also be interpreted as a measure of the semantic content of a passage. The average PDensity of each speakers' sentences was computed using the step-by-step procedures given in Turner and Greene (1977).

Reliability. One coder analyzed all 1,500 sentences; the sentences were randomized such that sentences from the same narrative or the same speaker were not analyzed consecutively. A second coder independently analyzed 100 sentences using 10 of the metrics; intercoder reliability was high for all these measures: MLU = 100%; MCU = 100%; DSS = 98%; IPSyn = 92%; DLevel = 100%; DComplexity = 94%; Maximal Yngve = 100%; Total Yngve = 100%; Local Frazier = 95%; Total Frazier = 94%. Ten

Table 1. *The results of the univariate ANOVAs for the 11 complexity metrics*

	Age group	Linear
	$F(2, 27)$	$F(1, 27)$
MLU	2.36	2.14
MCU	4.72*	4.36*
DSS	5.24*	4.38*
DLevel	15.87**	11.04**
IPSyn	1.68	1.12
DComplexity	4.89*	4.27*
MaximalY	11.98*	4.61**
TotalY	20.20**	13.78**
LocalF	5.51*	5.01*
TotalF	14.71*	12.37*
PDensity	2.03	1.01

$^{*}p < .05$; $^{**}p < .01$.

of the language samples had been previously propositionalized as part of a prior analysis of adults' narrative structure (Kemper et al., 1990); for this analysis, two coders independently analyzed each language sample, and intercoder agreement for PDensity was 94%. Split-half reliabilities for 10 of the measures were high, ranging from 92% for Total Frazier to 98% for DLevel; the split-half reliability for PDensity was somewhat lower, 85%.

Results

The 11 complexity measures were compared by performing a MANOVA with age group of the speaker as the between-subjects factor. The multivariate effect of age group was significant, $F(2, 270) = 19.83, p < .01$, and 8 measures produced significant age effects: MCU, DSS, DLevel, DComplexity, Maximal Yngve Depth, Total Yngve Depth, Local Frazier count, and Total Frazier count. The univariate *F*s are listed in Table 1. For these metrics, the linear component of the age effect was significant in each case; the higher order polynomial trends were not significant. No significant age group differences were found for the remaining 3 measures, MLU, IPSyn, and PDensity. Figure 5 plots age-group differences for each measure.

Table 2 presents the matrix produced by correlating the 11 complexity measures with the speakers' age, educational level, vocabulary, and digit span scores. Individual differences in complexity appear to reflect an age-related decline in working memory in producing sentences with multiple levels of embedding since speaker age was negatively correlated with MCU, DSS, DLevel, DComplexity, both Yngve Depth measures, and both Frazier counts, and since digit spans were positively correlated with these same

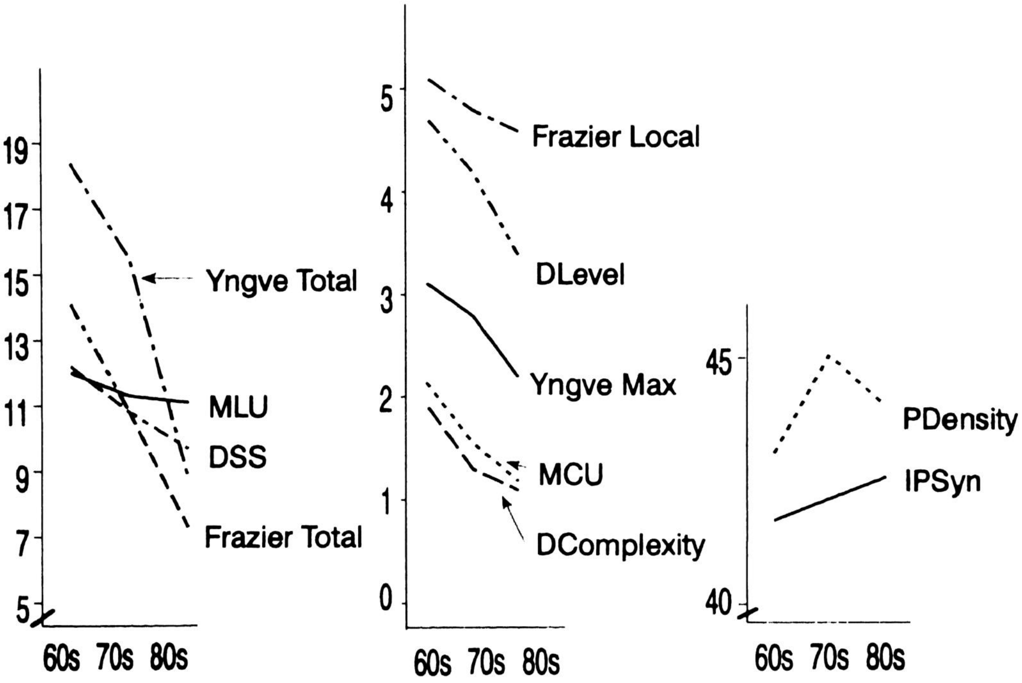

Figure 5. Age-group differences in complexity, according to each of the 11 metrics.

Table 2. *Matrix of correlations between the complexity measures and the speakers' age, educational level, vocabulary score, and digit span*

	Age	Education	Vocabulary	Digit span
MLU	−.07	+.30	+.28	+.04
MCU	−.52**	+.11	+.14	+.82**
DSS	−.46*	+.13	+.15	+.59**
DLevel	−.48*	+.13	+.16	+.61**
IPSyn	−.03	+.34	+.35	+.02
DComplexity	−.41*	+.11	+.17	+.52**
MaximalY	−.52**	+.14	+.17	+.66**
TotalY	−.54**	+.15	+.17	+.69**
LocalF	−.51**	+.13	+.17	+.65**
TotalF	−.58**	+.15	+.19	+.74**
PDensity	−.11	+.31	+.29	−.06

*p < .05; **p < .01, two-tailed.

metrics. Older speakers produced sentences with fewer embedded clauses, thus lowering all these measures. Speakers with greater digit spans produced sentences that contained more embedded clauses; this increased all these metrics. Individual differences in educational level or vocabulary are not correlated with these measures of complexity, although they are somewhat correlated with MLU and IPSyn. Better educated adults and adults with larger vocabularies tend to produce longer sentences, as measured by MLU, and sentences containing a greater variety of grammatical forms, as measured by IPSyn. PDensity is weakly correlated with educational level and vocabulary, suggesting that better educated adults and adults with larger vocabularies tend to pack more ideas into fewer words than other speakers. Table 3 presents the matrix of correlations among the 11 complexity measures.

To further examine age-group and individual differences in linguistic complexity as a function of length, the amount and type of embedding, and semantic content, structural equation modeling using EQS (Bentler, 1989) was used to test the fit of various models of linguistic complexity. The series of models was designed to clarify the relationship of amount of embedding, type of embedding, and semantic content as alternative sources of linguistic complexity.

The input to the structural equation models was a variance–covariance matrix including: speaker age, WAIS vocabulary, WAIS digit span (summed forward and backward span), educational level, and the average score on each of the 11 complexity metrics. Each structural model was evaluated using the maximum likelihood chi-square approach which measures the goodness-of-fit of the covariance matrix predicted by the model to the observed, input matrix. EQS fits both a measurement model, including factor loadings and measurement errors, and a structural equation model of the relations among the endogenous or independent variables (age, vocabulary, digit span, and education) and the dependent variables (the complexity metrics), which define latent factors. The fit of a series of structural equation models was then tested against the input covariance matrix. The series of models specified different latent factors, measured by various combinations of the complexity metrics.

The first model to be tested is summarized in Figure 6. In this model, Linguistic Complexity was assumed to be a single dependent latent factor which was measured by 11 metrics; Verbal Ability$_{\text{Age}}$ was also assumed to be an independent latent variable which reflected the common variation among age, educational level, vocabulary, and digit span; in other words, this variable is the age-related component of verbal ability as measured by education, vocabulary, and digit span. This model specified that Verbal Ability$_{\text{Age}}$ determines Complexity. A significant chi-square, $\chi^2(90) = 1036.90, p < .001$, was obtained. This indicates that the model does not fit the observed variance–covariance matrix and can be rejected.

A series of further models were then specified by defining additional latent factors by measured variables and the paths holding between the latent factors. The goal was to find a model, using maximum likelihood estimation procedures, that fits the data as well as the saturated model in which each measured variable corresponds to a latent factor and each factor is correlated with all the other factors. The series of models is summarized in Table 4.

The second model tested differed from the first in specifying two corre-

Table 3. *Matrix of correlations among the 11 complexity measures*

	MLU	MCU	DSS	DLevel	IPSyn	DComplex	MaxY	TotalY	LocalF	TotalF	PDensity
MLU	–										
MCU	.04	–									
DSS	.33	.66**	–								
DLevel	.34	.69**	.63**	–							
IPSyn	.29	.02	.02	.02	–						
DComplex	.02	.59**	.54**	.56**	.02	–					
MaximalY	.36	.74**	.68**	.70**	.02	.60**	–				
TotalY	.38*	.77**	.71**	.73**	.02	.63**	.79**	–			
LocalF	.36	.73**	.67**	.69**	.02	.60**	.74**	.78**	–		
TotalF	.41*	.83**	.76**	.79**	.02	.68**	.84**	.89**	.84**	–	
PDensity	–.05	–.06	–.06	–.07	.20	–.06	.07	–.07	.07	–.08	–

*$p < .05$; **$p < .01$, two-tailed.

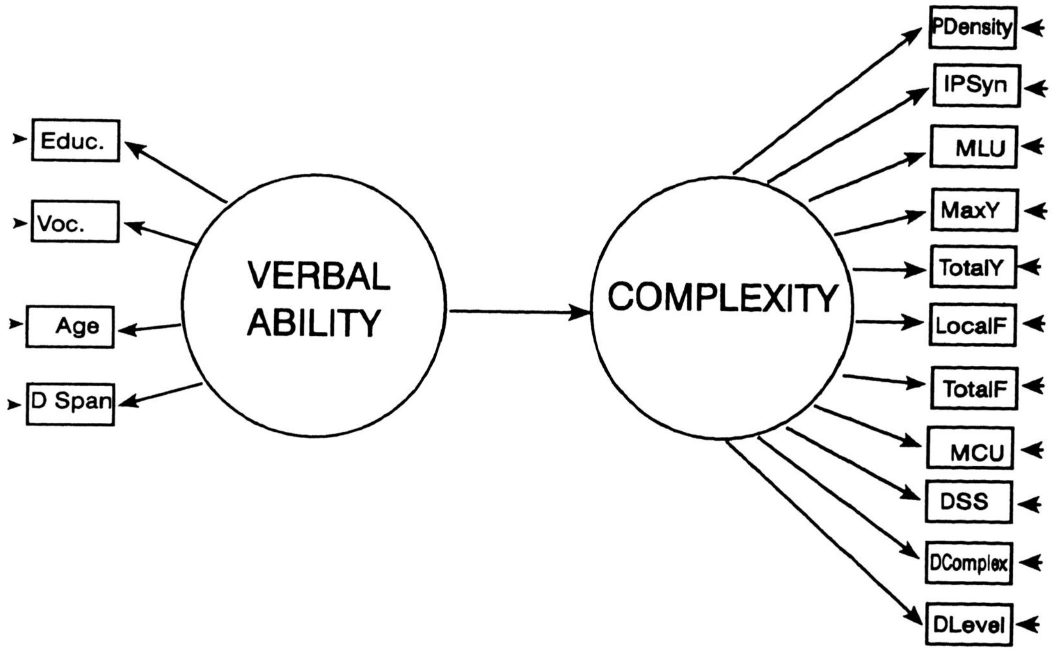

Figure 6. Initial structural equation model specifying one independent factor, Verbal Ability, and one dependent factor, Complexity.

Table 4. *Summary of structural equation models that were tested against the data*

Independent factors	Dependent factors	χ^2	*df*	*p*
1. Verbal ability	Complexity	1036.90	90	<.001
2. Verbal ability Working memory	Complexity	960.15	85	<.001
3. Verbal ability Working memory	Length Complexity	769.98	82	<.001
4. Verbal ability Working memory	Content Length Embedding Complexity	36.98	80	>.50
5. Verbal ability Working memory	Content Length Amount of Embedding Type of Embedding Complexity	15.09	79	>.50

lated factors, Verbal $\text{Ability}_{\text{Age}}$, the age-related change in verbal ability as measured by education and vocabulary, and Working $\text{Memory}_{\text{Age}}$, the age-related change in working memory as measured by digit span. One dependent factor, Complexity, was specified as jointly determined by Verbal $\text{Ability}_{\text{Age}}$ and Working $\text{Memory}_{\text{Age}}$. This model provided a closer ap-

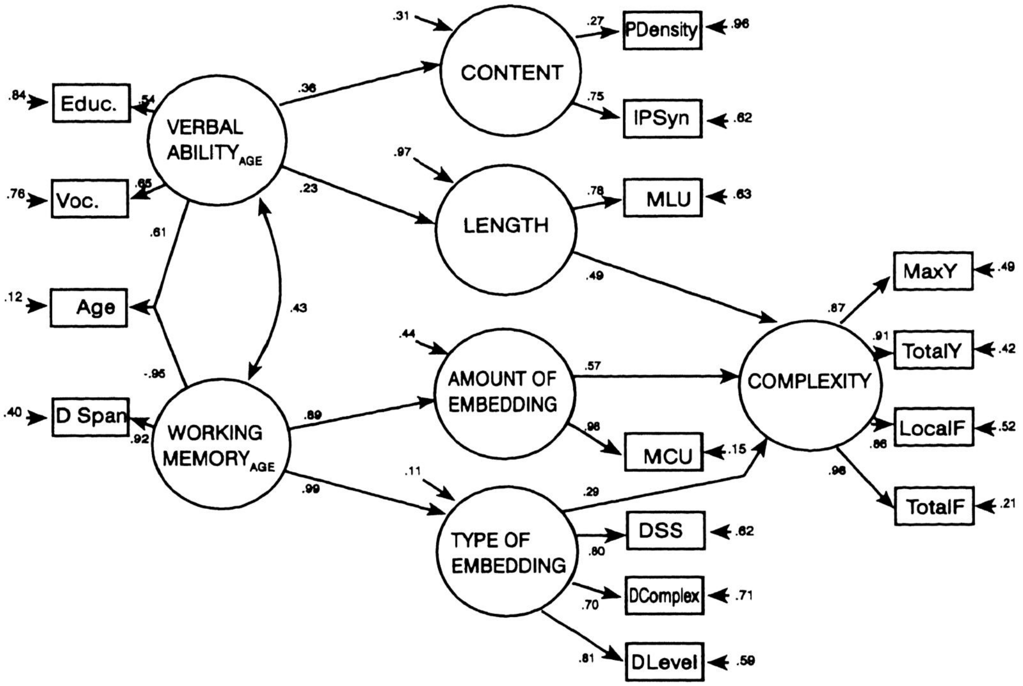

Figure 7. Final model specifying two correlated, independent factors, Verbal Ability and Working Memory; four dependent factors, Length, Amount of Embedding, and Type of Embedding, which determine Complexity; and a fifth factor, Content, which is unrelated to Complexity.

proximation to the data, but it still does not reproduce the observed variance–covariance matrix.

The third model specified both Verbal Ability$_{Age}$ and Working Memory$_{Age}$ as correlated factors, as in model 2, and distinguished two dependent factors, Length measured by MLU and Complexity measured by the remaining 10 metrics. Both Length and Complexity were specified as determined jointly by Verbal Ability$_{Age}$ and Working Memory$_{Age}$. This model, therefore, tested whether there are differences in linguistic complexity due to age-related changes in the speakers' verbal ability and working memory apart from those associated with sentence length. This model also does not fit the data.

The fourth model assumed two correlated factors, Verbal Ability$_{Age}$ and Working Memory$_{Age}$, and three dependent factors: Length, measured by MLU; the Amount of Embedding, measured by MCU; and Complexity, measured by DSS, DLevel, DComplexity, both Yngve metrics, and both Frazier metrics. PDensity and IPSyn were specified as loading on a Content factor which was predicted by the Verbal Ability$_{Age}$ factor. Length was also predicted by Verbal Ability$_{Age}$, whereas Embedding was specified as determined by Working Memory$_{Age}$. This model provides a close approximation to the data.

The fifth and final model is summarized in Figure 7; this model fits the data by specifying two correlated factors, Verbal Ability$_{Age}$ and Working

Memory$_{Age}$. Age is negatively associated with Working Memory, leading to a decline in digit span with advancing age, and somewhat positively associated with Verbal Ability, reflecting a slight improvement in vocabulary with advancing age. Linguistic Complexity is determined by three latent factors: Length, measured by MLU; the Amount of Embedding, measured by MCU; and the Type of Embedding, measured by DSS, DLevel, and DComplexity. Linguistic Complexity, in this model, is measured by both Yngve Depth metrics and both Frazier counts. Length, the Amount of Embedding, and the Type of Embedding are not correlated over and above their intercorrelation with Verbal Ability$_{Age}$ and Working Memory$_{Age}$. However, Length, the Amount of Embedding, and the Type of Embedding contribute to linguistic Complexity. Finally, Verbal Ability$_{Age}$ predicts another latent factor, Content, measured by both PDensity and IPSyn, which is not correlated with Complexity. Model 5 fits the data, $\chi^2(79) = 15.09$, $p > .50$, comparative fit index = .944.

Discussion

With the exception of PDensity and IPSyn, each of the complexity metrics appears to provide an adequate account of age-group and individual differences in linguistic complexity. These metrics, MLU, MCU, DSS, DLevel, DComplexity, Maximal and Total Yngve depth, and Local and Total Frazier node count, are sensitive to the effects of age, verbal ability, and working memory on the production of complexity grammatical constructions. The structural equation models indicate that verbal ability and working memory are correlated factors that change with advancing age and determine how speakers' sentence length, the amount of embedding, and the type of embedding vary with advancing age. The final model also indicates that these three factors - Length, Amount of Embedding, and Type of Embedding - determine the overall complexity of adults' speech independently of the semantic content of speech, as measured by PDensity, and the grammatical content of speech, as measured by IPSyn.[2]

EXPERIMENT 2

Experiment 2 was designed to determine whether syntactic complexity determines sentence comprehensibility and to evaluate each metric as a measure of sentence comprehensibility. Sentences varying in complexity were used as stimuli. A panel of judges listened to the sentences against a background of white noise, rated the sentences' comprehensibility, and then attempted to recall the sentences verbatim.

Method

Subjects. Five graduate students in speech-language-hearing or related fields served as judges. Each was naive with respect to the purposes of the study or the source of the stimuli.

Table 5. *Matrix of correlations between ratings of comprehensibility and verbatim recall and the complexity measures*

	Comprehensibility	Verbatim recall
	$r(98)$ =	$r(98)$ =
MLU	−.43**	−.47**
MCU	−.60**	−.61**
DSS	−.54**	−.48**
DLevel	−.59**	−.51**
IPSyn	−.12	+.14
DComplexity	−.54**	−.53**
MaximalY	−.64**	−.61**
TotalY	−.50**	−.69**
LocalF	−.62**	−.58**
TotalF	−.61**	−.67**
PDensity	+.21*	+.19

*$p < .05$; **$p < .01$, two-tailed.

Stimuli. One hundred sentences were selected from the language samples analyzed by Kemper et al. (1990). Five sentences were selected from each of 20 different speakers. These sentences were analyzed by both the primary and the secondary coder and scored on each of the 11 complexity metrics; intercoder reliability was better than 95% for each metric. The sentences were audiorecorded, in a random order, by a female speaker. The recording was then mixed with white noise. The sentences were presented binaurally over speakers in an audiometric room at 75 db with a −15 db signal-to-noise ratio. A 30-second pause, filled by white noise, occurred after each sentence.

Procedure. The judges were tested simultaneously. The judges received test booklets in which to record their responses, and rated each sentence on a 10-point scale ranging from (1) *very easy to understand* to (10) *very difficult to understand.* After rating each sentence, the judges then attempted to write down the sentence verbatim.

Results

Two measures were obtained for each sentence: the mean comprehensibility rating, averaged over the five judges, and the mean proportion of words, recalled by the five judges. These measures were somewhat correlated, $r(3) = +.58, p > .10$, indicating that they were not independent responses.

The comprehensibility ratings and recall scores were then correlated with the 11 measures of syntactic complexity obtained in Experiment 1. The matrix of correlations is presented in Table 5. Comprehensibility and recall were negatively correlated with complexity, with the exception of the IPSyn and PDensity metrics; this indicates that longer sentences, as measured by MLU, and more complex sentences, as measured by MCU, DSS, DLevel, DComplexity, both Yngve metrics, and both Frazier metrics, were more difficult to understand and recall than less complex sentences. Variations in IPSyn and PDensity do not appear to affect sentence comprehensibility and verbatim recall.

Hierarchical multiple regression was then used to identify the best set of complexity metrics for predicting sentence comprehensibility and recall. In the analyses, a hierarchical procedure was used in which sets of variables were sequentially added to regression equations for predicting comprehensibility or verbatim recall; the steps were ordered to reflect the structural equation model obtained in Experiment 1. At each step, the best predictor of a set of one or more intercorrelated variables, defining a latent factor, was entered into the regression equation. At each step, only those variables were entered into the equation whose partial correlation with comprehensibility and recall (with the effects of all previously entered variables controlled) was significant at $p < .05$ or better for the F-to-enter statistic. Then, at each step, any improvement in prediction of the resulting regression equation reflects the contribution of that step after the effects of all previously entered variables have been partialed out. Hierarchical multiple regression was used to avoid problems associated with multicollinearity since these measures of syntactic complexity are highly intercorrelated.

In step 1, semantic Content, measured by PDensity and IPSyn, was initially used to predict comprehensibility and recall. In step 2, sentence Length, measured by MLU, was added. In step 3, the Amount of Embedding, measured by MCU, was entered into the regression model. In step 4, Type of Embedding, measured by DSS, DLevel, and DComplexity, was added. In step 5, overall Complexity, measured by both Yngve and both Frazier metrics, was added. The results are summarized in Table 6.

In these analyses, PDensity, DLevel, and Maximal Yngve Depth were selected as the best predictors of Content, Type of Embedding, and overall Complexity, respectively. As Table 6 indicates, comprehensibility and recall are not predicted by semantic Content, and adding sentence Length to Content results in a marginally insignificant improvement in the fit of the regression equation. Adding the Amount of Embedding does significantly improve the prediction of both comprehensibility and recall, and a further significant improvement is obtained by adding the Type of Embedding. None of the overall measures of Complexity resulted in a further improvement in the fit of the regression model for either comprehensibility or recall.

Discussion

Experiment 2 validates the model of linguistic complexity that resulted from the language sample analysis conducted in Experiment 1. Sentence Length, Amount of Embedding, and Type of Embedding not only determine the

Table 6. *Results of the hierarchical multiple regression analyses of comprehensibility and verbatim recall*

		Comprehensibility		Recall	
		R^2	*F*(change)	R^2	*F*(change)
1	Content				
	PDensity	.19	–	.15	–
2	Length				
	MLU	.27	3.75*	.29	3.98*
3	Amount of Embedding				
	MCU	.39	9.87**	.43	11.23**
4	Type of Embedding				
	DLevel	.56	9.24**	.57	10.05**
5	DComplexity				
	MaximalY	.61	1.21	.59	<1.0

*$p < .05$; **$p < .01$.

complexity of adults' speech but also determine how easily individual sentences can be understood and how accurately individual sentences can be recalled.

GENERAL DISCUSSION

Experiment 1 confirmed prior research by showing that the complexity of adults' speech declines with advancing age and appears to reflect a reduction in the capacity of working memory rather than differences in education or vocabulary. The complexity of adults' speech declined with age as measured by 8 of the 11 metrics: MCU, DSS, DLevel, DComplexity, both Yngve Depth Measures, and both Frazier counts. Semantic content does not appear to vary with age, nor does MLU in words or the IPSyn inventory of grammatical forms.

The structural equation modeling suggests that the age-related decline in complexity occurs because of a reduction in sentence length and a loss of embedding per se as well as a loss of particular types of sentence embeddings, such as left-branching embeddings. Sentence embeddings impose demands on working memory for the simultaneous construction and manipulation of multiple syntactic constituents. Since working memory appears to decline with advancing age, adults become less able to construct complex syntactic structures with embedded gerunds, *that* and *wh-* clauses, and other embedded and subordinate structures.

Sentence complexity is somewhat determined by sentence length, as measured by MLU, as well as by the amount of embedding and subordination, as measured by MCU, and the type of embedding and subordination, as measured by DSS, DLevel, and DComplexity. Embeddings, such as the use of *that* and *wh-* clauses as sentential subjects (particularly those that produce left-branching structures), increase DSS, DLevel, and DComplexity. Thus, not only does the amount of embedding per se increase linguistic complexity, but the type of embedding also contributes to complexity. Linguistic complexity can be measured directly by computing either Yngve Depth metric or either Frazier count. These measures are sensitive to variation in length, amount of embedding, and type of embedding.

This model of linguistic complexity was validated by Experiment 2. How easily sentences are understood and how accurately they are recalled cannot be predicted on the basis of the content and length of sentences as measured by PDensity and MLU, respectively. Rather, sentence comprehensibility and recall reflect the amount of embedding, as measured by MCU, and type of embedding, as measured by DLevel. The resulting regression formula with four predictors - content, length, amount of embedding, and type of embedding - accounts for 78% of the variance in comprehensibility and 76% of the variance in recall.

Applications

In looking for the determinants of sentence processing difficulties, psycholinguists have identified many contributing syntactic factors either by systematically contrasting sentences with different syntactic properties, or by developing formulae for ordering sentences as to their overall complexity. The choice of a complexity metric for research purposes will depend on practical considerations. For most language samples, MLU and MCU can be easily computed; however, MLU, while widely used to scale children's language acquisition, shows little variation over the adult years and may not be sensitive to developmental differences once the basics of morphology and syntax have been mastered (Kemper et al., 1989; Klee & Fitzgerald, 1985). MCU has limited utility for the study of the early stages of language acquisition, since young children do not begin to master the syntax of embedding until rather late in the acquisition period (Limber, 1973).

While MLU and MCU can be computed with some ease, the other complexity metrics require skilled analysis for their application. DSS, DLevel, and DComplexity require that the researcher carefully examine each sentence for a wide range of different syntactic constructions and assign appropriate point values to these constructions. The Yngve and Frazier metrics require the researcher to perform a surface phrase structure analysis of the sentence. The Frazier analysis is more difficult to execute than the Yngve analysis, since it attempts to emulate a deterministic, left-to-right parser. The analysis must detect and fill in gaps in the structure of the sentence whenever noun phrases have been deleted or fronted. For example, a gap occurs in "The students tried to learn" since the subject of the infinitive ("the students"), which is coreferential with the subject of the main clause, has been deleted. This gap contributes to the complexity of the sentence according to the Frazier analysis, although it makes no contribution to any of the other analyses.

The most difficult metric to compute is PDensity. The reliable identification of individual propositions requires extensive training. The process of propositionalizing an entire language sample is also slow; for adult speakers, PDensity averages approximately 43 and can range from 20 to 80 (Kemper et al., 1990). Thus, for a sample of 100 words, between 20 and 80 propositions may have to be identified.

For these reasons, MCU may provide an adequate index of linguistic complexity for many purposes; this measure can be easily calculated, it appears to be a central determinate of age-related and individual differences in linguistic complexity, and it correlates strongly with comprehensibility and verbatim recall. Additional control over linguistic complexity can be gained by computing DLevel.

ACKNOWLEDGMENTS

This research was supported by grants RO1AG06319 and KO4AG00443 from the National Institute on Aging to S. Kemper and a Biomedical Research Award to the University of Kansas. We would like to thank Donna Kynette, Suzanne Norman, and Shannon Rash for their assistance with collecting and transcribing the language samples, Cliff Pye for his advice on matters linguistic, and Susan Embretson and David Thissen for their advice on matters statistical.

NOTES

1. Working memory limitations have been implicated in some forms of childhood reading impairments (Gathercole & Baddeley, 1989, 1990; Shankweiler & Crain, 1986) and individual differences in reading comprehension (Daneman & Carpenter, 1980). Baddeley (1986) proposed a tripartite model of working memory in which a central executive component is responsible for most processing operations and two subordinate storage systems, an articulatory loop and a visual-spatial sketchpad, provide temporary storage of verbal and visual-spatial information, respectively. Under this framework, processing deficits can arise because the central executor is overloaded with processing operations or because the capacity of either temporary store is exceeded. Daneman and Tardiff (1987) suggested that the articulatory loop and visual-spatial sketchpad are not simply storage systems but also specialized, limited-capacity processors; the articulatory loop would correspond to the syntactic parser. Under this reformulation, working memory limitations on language processing would be expected to arise from limitations on the kinds of syntactic operations that can be performed by the parser during production and comprehension. Other conceptions of working memory (Hasher & Zacks, 1988; Salthouse, Babcock, & Shaw, 1991) also imply that age-related differences in working memory will hinder older adults' language processing.
2. The small sample size leads to some cautions in the interpretation of the present series of models: (a) the model parameter estimates may be sample-specific since the input covariance matrix may not provide asymptotic estimates of the population covariance matrix. (b) The likelihood ratio tests have low power and may lead to Type II errors such that a model will be accepted that would have been rejected on the basis of a larger sample size. Further, the input to

this series of models was a covariance matrix, rather than a correlation matrix; rescaled, standardized parameter estimates are reported in Figure 7 for the final model since standardized statistics are easier to interpret. Only the maximum likelihood estimates of the final model are reported. Readers interested in the maximum likelihood estimates for the rejected models should contact S. Kemper.

REFERENCES

Baddeley, A. (1986). *Working memory.* Oxford: Oxford University Press.

Bentler, P. M. (1989). *EQS: Structural Equations program manual.* Los Angeles: BMDP Statistical Software.

Berwick, R. C., & Weinberg, A. S. (1984). *The grammatical basis of linguistic performance: Language use and acquisition.* Cambridge, MA: MIT Press.

Botel, M., & Granowsky, A. (1972). A formula for measuring syntactic complexity: A directional effect. *Elementary Education, 49,* 513–516.

Bresnan, J. (Ed.). (1982). *The mental representation of grammatical relations.* Cambridge, MA: MIT Press.

Brown, R. (1973). *A first language.* Cambridge, MA: Harvard University Press.

Clancy, P. M., Lee, H., & Zoh, M. H. (1986). Processing strategies in the acquisition of relative clauses: Universal principles and language-specific realizations. *Cognition, 24,* 225–262.

Crain, S., & Schankweiler, D. (1988). Snytactic complexity and reading acquisition. In A. Davison & G. Green (Eds.), *Critical approaches to readability* (pp. 167–192). Hillsdale, NJ: Erlbaum.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Ability, 19,* 450–466.

Daneman, M., & Tardiff, T. (1987). Working memory and reading skill reexamined. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 491–508). Hillsdale, NJ: Erlbaum.

Fay, D. (1980). Performing transformations. In R. L. Cole (Ed.), *Perception and production of fluent speech* (pp. 441–468). Hillsdale, NJ: Erlbaum.

Fodor, J. A., Bever, T. G., & Garrett, M. (1974). *The psychology of language.* New York: McGraw-Hill.

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior, 22,* 203–218.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computation, and theoretical perspectives* (pp.129–189). Cambridge: Cambridge University Press.

(1988). The study of linguistic complexity. In A. Davison & G. Green (Eds.), *Critical approaches to readability* (pp. 193–223). Hillsdale, NJ: Erlbaum.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 6,* 291–325.

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28,* 200–213.

(1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29,* 336–360.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, 193–226). New York: Academic.

Hsu, J. R., Cairns, H. S., & Fiengo, R. W. (1985). The development of grammars underlying children's interpretation of complex sentences. *Cognition, 20,* 25–48.

Jackendoff, R. (1977). *Bar-$\bar{X}$ syntax: A study of phrase structure.* Cambridge, MA: MIT Press.

Kemper, S. (1986). Imitation of complex syntactic constructions by elderly adults. *Applied Psycholinguistics, 7,* 277-287.

(1987a). Life-span changes in syntactic complexity. *Journal of Gerontology, 42,* 323-328.

(1987b). Syntactic complexity and the recall of prose by middle-aged and elderly adults. *Experimental Aging Research, 13,* 47-52.

(1988). Geriatric psycholinguistics: Syntactic limitations of oral and written language. In L. Light & D. Burke (Eds.), *Language and memory in old age* (pp. 58-76). Cambridge: Cambridge University Press.

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

Kemper, S., & Rash, S. (1988). Speech and writing across the life-span. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (pp. 107-112). Chichester, U.K.: Wiley.

Kemper, S., Rash, S. R., Kynette, D., & Norman, S. (1990). Telling stories: The structure of adults' narratives. *European Journal of Cognitive Psychology, 2,* 205-228.

Kintsch, W., & Keenan, J. M. (1973). Reading rate and retention as a function of the number of the propositions in the base structure of sentences. *Cognitive Psychology, 5,* 257-274.

Kintsch, W., & Vipond, D., (1978). Reading comprehension and readability in educational practice and psychological theory. In L. G. Nilsson (Ed.), *Perspectives on memory research* (pp. 329-365). Hillsdale, NJ: Erlbaum.

Klee, T., & Fitzgerald, M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language, 12,* 251-269.

Kynette, D., & Kemper, S. (1986). Aging and the loss of grammatical forms: A cross-sectional study of language performance. *Language and Communication, 6,* 43-49.

Lee, L. (1974). *Developmental sentence analysis.* Evanston, IL: Northwestern University Press.

Limber, J. (1973). The genesis of complex ideas. In T. Moore (Ed.), *Cognitive psychology and the acquisition of language* (pp. 169-186). New York: Academic.

Marcus, M. (1980). *A theory of syntactic recognition for natural language.* Cambridge, MA: MIT Press.

McCawley, J. D. (1989). *The syntactic phenomena of English* (Vols. 1-2). Chicago: University of Chicago Press.

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

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

Salthouse, T. A., Babcock, R. L., & Shaw, R. J. (1991). Effects of adult age on structural and operational capacities in working memory. *Psychology and Aging, 6,* 118-128.

Scarborough, H. S. (1990). Index of Productive Syntax. *Applied Psycholinguistics, 11,* 1-22.

Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorder: A modular approach. *Cognition, 24,* 139-168.

Shapiro, L. P., Zurif, E., & Grimshaw, J. (1987). Sentence processing and the mental representation of verbs. *Cognition, 27,* 219-246.

Smith, C. S. (1988). Factors of linguistic complexity and performance. In A. Davison & G. Green (Eds.), *Critical approaches to readability* (pp. 247-280). Hillsdale, NJ: Erlbaum.

Smith, C. S., & van Kleeck, A. (1986). Linguistic complexity and performance. *Journal of Child Language, 13,* 398-408.

Turner, A., & Greene, E. (1977). *The construction and use of a propositional text base* (Technical Report). University of Colorado, Boulder.

Watt, W. C. (1970). On two hypotheses concerning psycholinguistics. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 137-220). New York: Wiley.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence.* Baltimore: Williams & Wilkins.

Yngve, V. (1960). A model and a hypothesis for language structure. *Proceedings of the American Philosophical Society, 104,* 444-466.