A System for Assessing Grammaticality in Connected Speech of Mildly Aphasic Individuals

Kathryn M. Yorkston
David R. Beukelman

Department of Rehabilitation Medicine
University of Washington, Seattle

The verbal performance of mildly aphasic individuals may be essentially error free on standardized tests, yet their deficits may be of social and vocational concern. Knowledge of the differences between high-level aphasic and normal speakers is of interest because it gives insight into the nature of language deficits and also allows for the establishment of appropriate assessment procedures.

The system for assessing grammaticality of connected speech samples which will be presented here is the outgrowth of previous work (Yorkston and Beukelman, 1977). Briefly, a picture description task was used to elicit samples of connected speech from normal and mildly aphasic speakers. These samples were quantified in terms of the number of concepts conveyed, the number of syllables used, and the time required. When one examines just the amount of information conveyed, mildly aphasic speakers do not differ from normal speakers. Figure 1 represents the means and one standard deviation of number of concepts produced by 48 normal adults, 30 normal geriatric speakers and 50 aphasic speakers. The aphasic speakers were classified according to severity of verbal deficit (percentiles on the verbal subtests of the PICA). Examination of Figure 1 reveals that there is an inverse relationship between severity and the number of concepts conveyed, but the two mildest groups of aphasic speakers did not differ from normal speakers in terms of the amount of information conveyed.

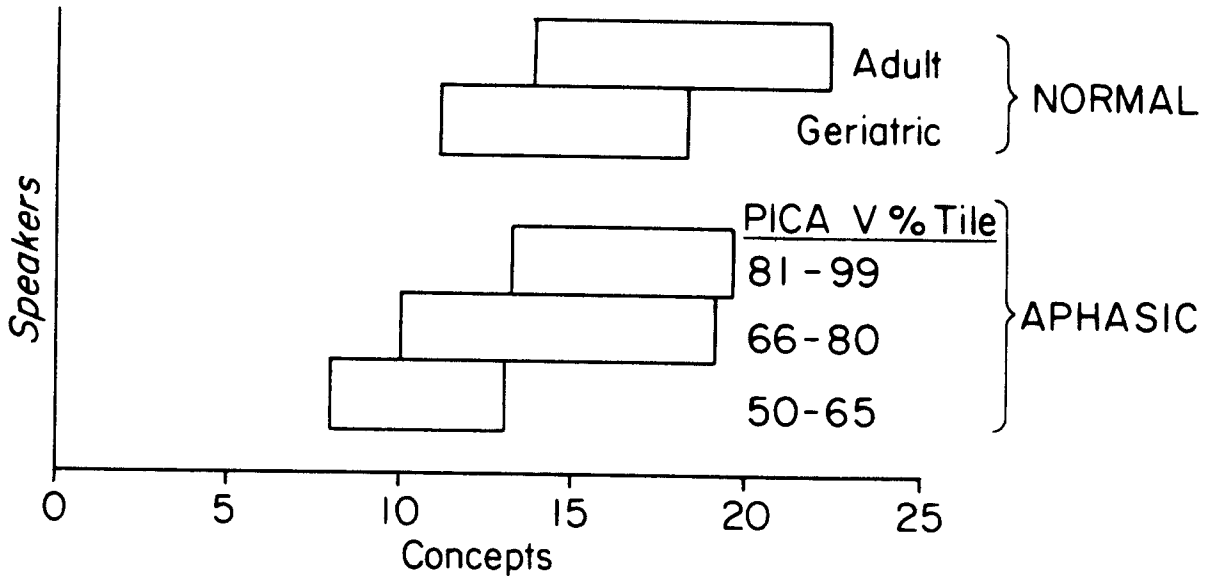

Figure 1. Bands indicating a one standard deviation range about the mean number of concepts communication by Normal Adults (N=48) and Geriatric (N=30) Speakers and Aphasic Speakers (N=50) classified according to verbal percentile on the PICA.

In order to differentiate performance of mildly aphasic speakers from normal speakers, time-based measures must be used. Figure 2 illustrates the mean and one standard deviation about the mean for syllables per minute and concepts per minute When efficiency or time-based measures are considered, there is little overlap between normal performance and that of high level aphasic speakers.

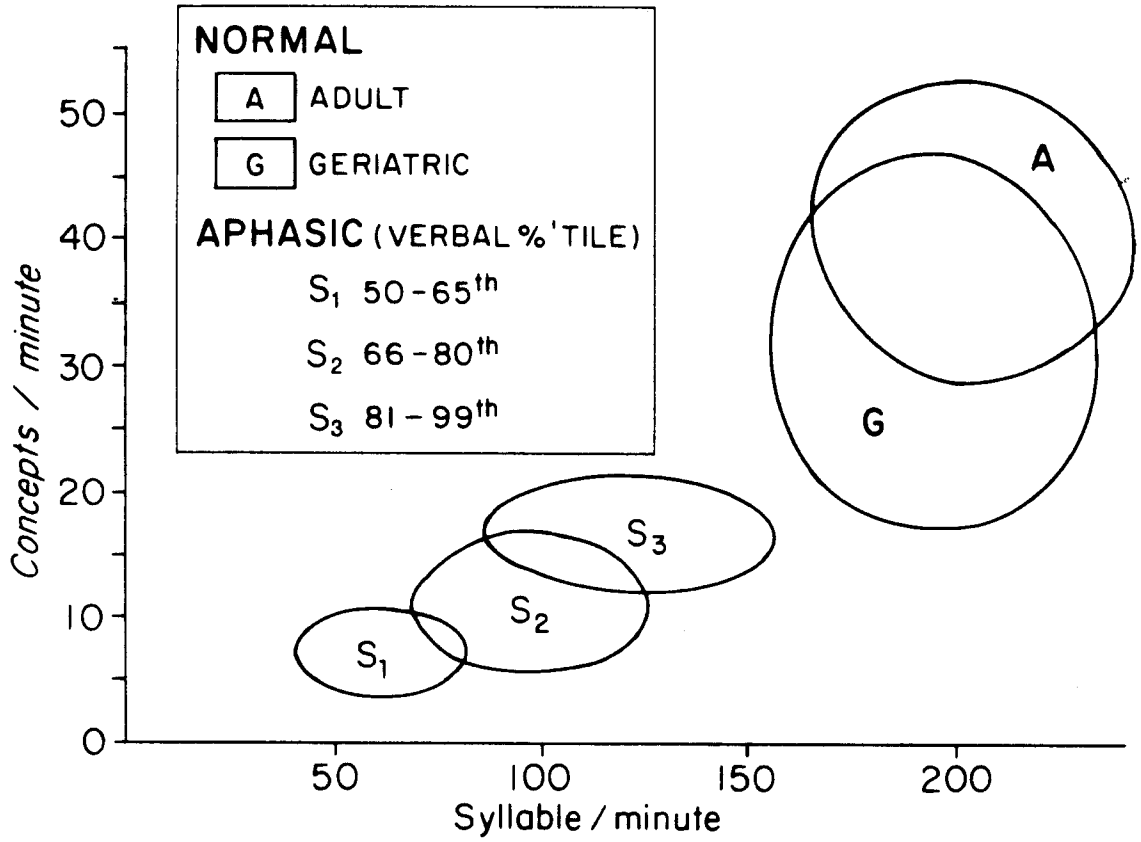

Figure 2. Means and one standard deviation about the means for syllables per minute and concepts per minute produced by Normal Adult and Geriatric Speakers and Aphasic Speakers.

Although the system developed was useful in quantifying the amount of information conveyed by high level aphasic speakers and the efficiency with which they communicated, it gave no indication of the grammatical structures being used or the grammatical errors being made. This study was an attempt to develop an index of grammaticality which can be applied to samples of connected speech of mildly aphasic speakers.

A review of the literature reveals that there are several methods for assessing grammaticality. Some researchers (Zurif and Caramazza, 1976) used structured tasks to elicit certain grammatical forms. However, some pilot work for the current project confirmed Goodglass's opinion (1976) that elicited samples do not accurately reflect the grammaticality of spontaneous speech.

More complex and a greater variety of forms could be elicited by asking questions designed to elicit specific grammatical constructions than by an unstructured picture description.

Howes and Geswind (1962) analyzed conversational samples by assessing dimensions such as number of personal pronouns and number of interstitial words. Their 5000-word samples were significantly larger than samples elicited by a picture description task. Therefore, this type of analysis was not used in the present studies.

Still another index of grammaticality which has been applied to children's language development is MEAN LENGTH OF UTTERANCE. Pilot work revealed several major problems when applying mean length of utterance measures to samples of connected speech of mildly aphasic individuals. The first is that, unlike conversational or spontaneous speech samples, it is virtually impossible in a picture description task to identify where one utterance ends and another begins. The second is that mildly aphasic speakers tend to use many parenthetical expressions within utterances. These parenthetical expressions are usually revisions or corrections, and, although they are indicative of a reduction in efficiency, they tend to artificially inflate the utterance length.

The system proposed in this paper is an index of grammaticality based on the mean length of uninterrupted grammatical strings. A String is defined as a series of words which have a grammatical relationship to each other. A String need not contain elements of sentence structure and may be a single word if none of the adjacent words are grammatically related. Only intelligible words are counted. A String is broken if any of the following occur:

> Two adjacent words are not grammatically related or there is a grammatical error, e.g. [the boy] [are stealing cookies].
>
> Falling intonation pattern clearly indicates the end of an utterance.
>
> Just prior to the second "and" of a compound sentence, e.g. [the girl is reaching up and the boy is falling] [and the woman is washing dishes.]

Appendix 1 contains a more complete description of the rules for identifying Strings and examples of utterances broken into grammatical Strings.

The speech samples from which the Yorkston and Beukelman (1978) data were derived were re-analyzed for grammaticality. The mean length of uninterrupted grammatical strings (MSL) was computed for each sample. The left side of Figure 3 illustrates the mean score for MSL and a one standard deviation band about the mean for normal and three aphasic severity groups. Examination of the figure reveals that there is an inverse relationship between severity and MSL, in that as severity decreases MSL increases. Although there is some overlap between the normal group and the highest group of aphasic speakers, statistics show that MSL for S_3 is lower than for normal speakers ($p. < .01$).

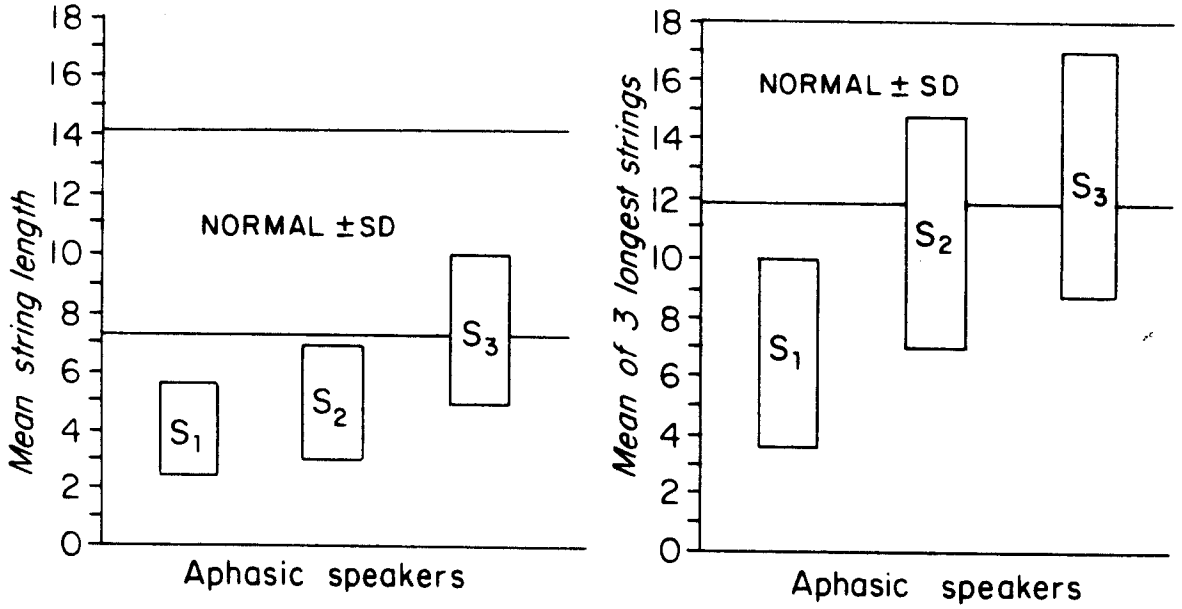

Figure 3. Mean String Length and Mean of the Three Longest Strings for Normal and Aphasic Speaker Groups S_1, S_2 and S_3.

These results seem to indicate that mildly aphasic speakers are not within normal limits in terms of grammaticality. However, a review of the transcripts of the picture description task revealed that grammatical strings were often broken by errors not specifically grammatical in nature. For example, the following excerpt is divided into three strings: [he is standing on a chair] [no that's not a chair it's a] [that's a stool]. The relatively brief string length here may not be the result of lack of grammatical knowledge or inability to apply that knowledge, but rather a word selection problem that interferes with the grammatical form of the final product. Because MSL assesses only the grammatical surface or the final grammatical product, it is sensitive to this type of reduction of efficiency.

In order to get a sample of "best performance" which would be as free as possible from these word selection interruptions, we also computed a measure of Mean of the Three Longest Strings (M3LS) for each speech sample. Means of this measure for each speaking group are plotted on the right of Figure 3. Comparison of the two measures of grammaticality shows that, as one would expect, scores of all the speaker groups tend to increase when only the best performance is considered. However, the scores of the aphasic speakers increase proportionately more than do the scores of the normal speakers. In fact, statistics show that speakers in S_3, those over the 80th percentile, are not significantly different from normal speakers.

Figures 4, 5 and 6 illustrate serial data obtained from aphasic speakers during treatment. The recovery graphs of three speakers will illustrate how MSL and M3LS change over time. Figure 4 represents measures of grammaticality obtained from a non-fluent speaker who moved from the 50th to the 79th percentile over a four-month period. Verbal output of this speaker could be characterized as non-fluent with little variety or variability in grammatical structure. Although grammatical forms were rather stereotyped, the grammatical flow was not interrupted by word selection problems. In terms of the measures of grammaticality proposed here, two points emerge. First, the best performance (M3LS) is reduced as compared to normal. Second, the best performance and the mean performance are very similar. Taken together, these measures suggest that grammatical performance is both less diversified and less complex than normal.

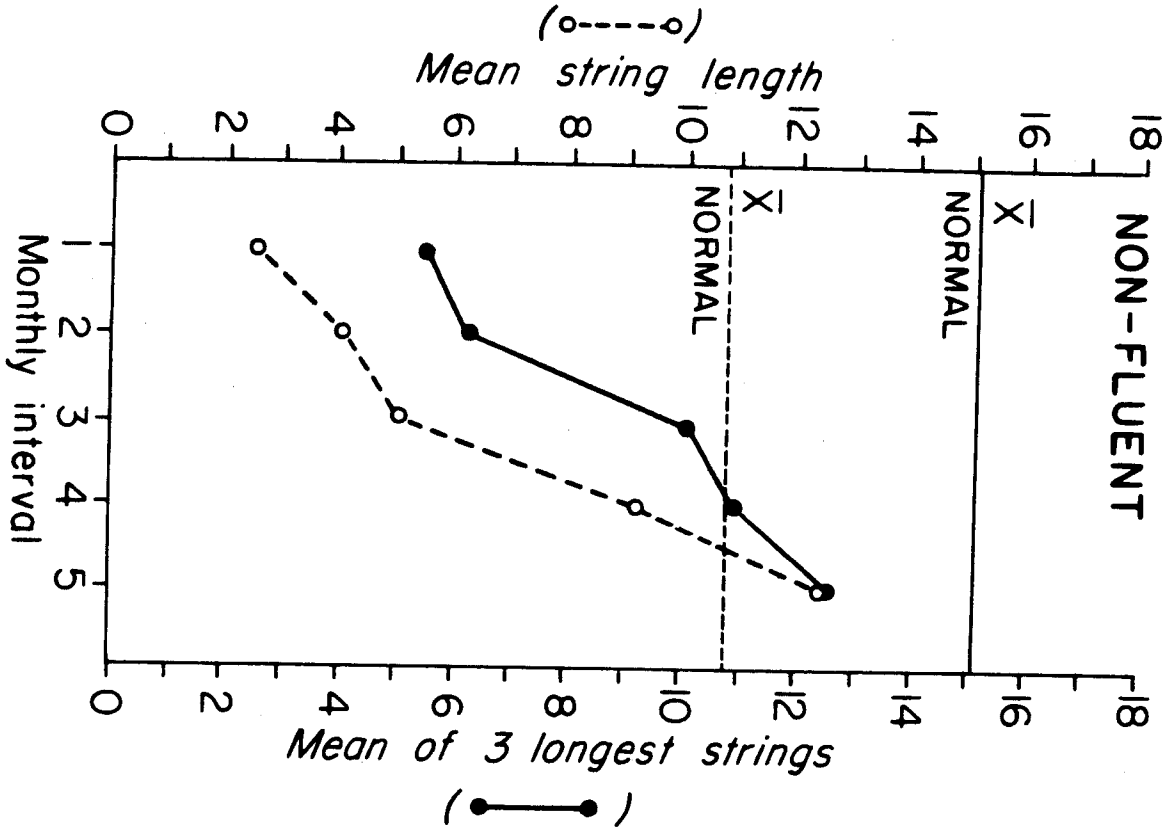

Figure 4. Mean String Length and Mean of Three Longest Strings for a Non-Fluent Aphasic Speaker during a four month period of treatment.

Figure 5 illustrates the recovery curve of a conduction aphasic speaker who moved from the 58th to the 90th percentile during the data collection period. His speech was fluent and during the initial months was marked by frequent articulatory breakdowns which he would attempt to self-correct. By the end of the data collection period, these articulatory breakdowns had essentially been eliminated and the most predominant remaining

deficit was an occasional word-finding problem. An examination of the measures of grammaticality revealed that by monthly interval "5," MSL was within normal limits and M3LS was over one standard deviation above normal. An interpretation of these data might be that this patient is capable of generating complex structures but that his overall performance does not match his best performance because of word selection problems.

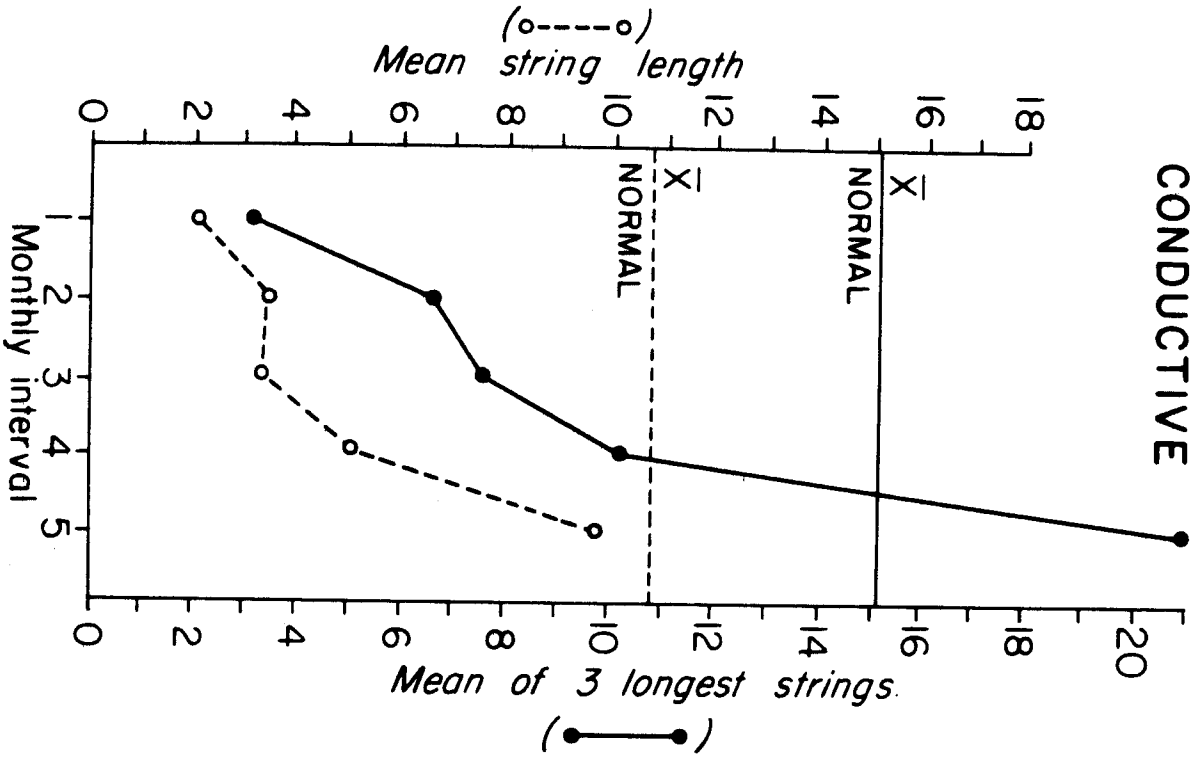

Figure 5. Mean String Length and Mean of Three Longest Strings for a conduction aphasic speaker during a four month period of treatment.

Some consistent patterns are beginning to emerge as more samples are collected. One of these patterns is that large gaps between best performance and overall performance usually indicate the presence of word selection problems. Figure 6 illustrates data obtained from an anomic speaker as he moved from the 55th to the 90th percentile. His speech was fluent, and throughout the course of recovery a word-finding problem was his most significant deficit. Examination of the measures of grammaticality revealed that by the third monthly interval this speaker was capable of generating grammatical strings as long as normal speakers. However, the gap between MSL and M3LS indicates that best performance was not matched by overall performance. In this case, the wide gap is due to the fact that grammatical flow was interrupted by frequent word selection problems.

In summary, the current study suggests that mildly aphasic speakers are capable of generating grammatical strings as long as strings generated by normal speakers. Despite the fact that their best performance is equivalent to the best performance of normal speakers, their overall performance does not match the overall performance of normal speakers.

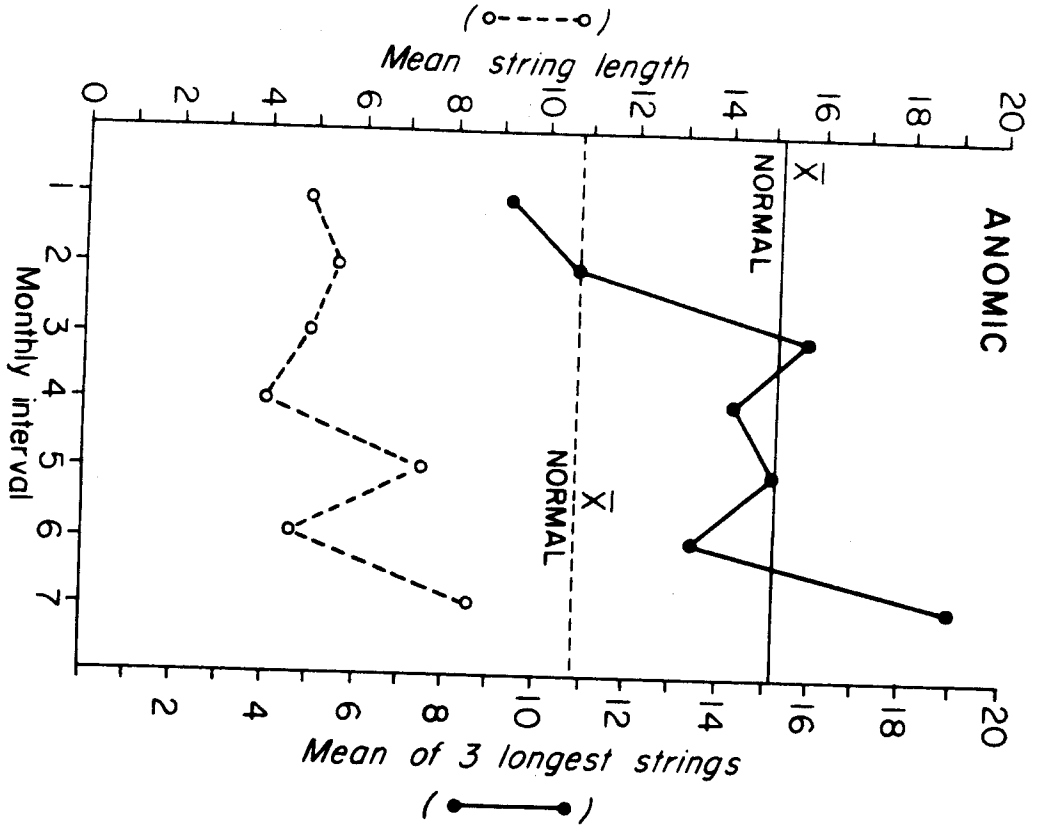

Figure 6. Mean String Length and Mean of Three Longest Strings for an Anomic Aphasic Speaker during a six month period of treatment.

This gap between best performance and overall performance indicates reduction in efficiency. However, because the measures of grammaticality proposed here look only at the grammatical form of the final product, it is impossible to say whether this reduction in efficiency is due to reduced grammatical knowledge, imprecise application of that knowledge, or other problems which are not grammatical in nature, such as word selection problems. Our previous work demonstrated that mildly aphasic speakers were able to convey as much information as normal speakers but they weren't able to do so as efficiently as normals, because the rate at which they conveyed information was slower than normal. These results, along with the results of earlier work suggest that mildly aphasic speakers are similar to normal speakers in their capacity to convey information and in their capacity to use long grammatical strings. However, their efficiency is slightly reduced in both of these areas.

APPENDIX I

Rules for Determining Mean Length of String and Mean of Three Longest Strings

I. Identification of Strings:

A verbal description of the "Cookie Theft" picture (Boston Diagnostic

transcribed. Note is made on the transcript of falling intonation or pauses which can be interpreted as meaning the end of an utterance. The following rules determine the beginning and end of a string.

1. Introductory words as phrases (e.g. "okay," "well," "here goes," "now let me see") and concluding words or phrases (e.g. "I guess that's it." "that's about all." "What else do you want?") are not considered as a STRING or a part of a STRING.
2. A grammatical STRING is broken if two adjacent words have no grammatical relationship to each other. Examples:
   a. [He's standing on a] -- [You know I just can't find that word] -- [on a stool.]
   b. [The boy and girl are trying to steal some cookies and their mother does not see them.] [The stool is falling.]
3. A STRING is broken if there is a grammatical error. Examples:
   a. [The boy] [reaching a cookie]
   b. [The boy are] [reaching a cookie]
4. A STRING is broken if there is a falling intonation pattern or a pause which can be interpreted as indicating the end of an utterance.
5. In the absence of a prosodic pattern marking the end of an utterance a STRING is broken just prior to the second "and" when a series of independent clauses are connected with "and."
6. When a speaker says several words in an attempt to produce the target word, the first attempt is considered part of the previous STRING. However, each subsequent word should be counted as a separate STRING. Example:
   a. [The boy is on the bed.] [Chair] [couch] [stool.]
7. A STRING remains unbroken when a speaker engages in a succession of attempts to correctly articulate a target word. None of the unsuccessful attempts should be counted as words.

Example:

   a. [The boy is on the s, st, sto, stool.]

8. A STRING is not broken by the presence of a "filler" sound ("uh") or word (okay).

Example:

   a. [The boy is falling off the.../uh/,okay /uh/...stool.]

9. A STRING is not broken by appositives.

Example:

   a. [There are two kids, a boy and a girl, who are stealing cookies.]

II. Counting Words:

After a response has been segmented into STRINGS, the total number of words is counted. The following rules apply:

1. A word must be intelligible.
2. Contractions are counted as two words.
3. If two STRINGS are identical, only the words in the initial STRING are counted,

   [The water is overflowing] [The water is overflowing].

   2 STRINGS
   4 Words

4. If the second string contains words not in the first, all of the words are counted.

   [The water is overflowing] [The water is overflowing onto the floor].

   2 STRINGS, 11 Words

III. Analysis:

4.1 Total number of words -- the total number of intelligible words in a response which are located in STRINGS.

4.2 The total number of STRINGS -- the total number of STRINGS within a response.

4.3 Mean String Length equals the total number of words divided by the total number of STRINGS.

$$\text{Formula: mean STRING length} = \frac{\text{total \# of words}}{\text{total \# of strings}}$$

4.4 Mean of three longest strings equals the total number of words in the three longest strings divided by three.

IV. Examples:

The following examples are excerpts taken from three different aphasic speakers. Strings are identified with brackets, the numbers above each string indicate the number of words in each string, and slashes through syllables indicate that the syllable has not been counted as a word.

[Kitchen]^{1} . . . [Lady's washing the dishes]^{5} . . . UH . . . [Water's running over]^{4} . . . [Girl]^{1} . . . [Cookie]^{1} . . . [Boy and Girl]^{3} . . . UH . . . [Wobbly]^{1} . . . [Outside]^{1} . . . [Nice Garden]^{2}.

WORDS = 19
STRINGS = 9
MEAN STRING LENGTH (MSL) = 2.1
MEAN of 3 LONGEST STRINGS (M3LS) = 4.0

[That woman really is . . . UH . . . RU, rubbing her dishes]^{7}. [The,]^{1} [She forgot to turn off the,] [Her water so the sink]^{7} is overflowing.] [The children]^{2} [they want a cookie]^{8} so the man is] [Lay,]^{1} [The boy is ST, ST,]^{8} stealing from the cookie jar.]

TOTAL WORDS: 40
TOTAL STRINGS: 8
MEAN STRING LENGTH (MSL): 5.0
MEAN OF 3 LONGEST STRINGS (M3LS): 7.7

[There's a lady there]^{11} and she's washing the dishes.] [And she's not aware of the kids trying]^{12} to get some carrots,] [Cookies and falling off the chair]^{6} [And the water's flowing]^{16} out of the sink and she's not aware of that either.]

TOTAL WORDS: 45
TOTAL STRINGS: 4
MEAN STRING LENGTH (MSL): 11.3
MEAN OF 3 LONGEST STRINGS (M3LS): 13.0

References

Goodglass, H., Agrammatism. In H. Whitaker and H.A. Whitaker (Eds.) Studies in Neurolinguistics, Vol.1, New York: Academic Press, 237-259 (1972).

Goodglass, H. and Kaplan, E. Boston Diagnostic Aphasia Examination. Philadelphia: Lea and Febiger (1972).

Howes, D. and Geshwind, N. Quantitative Studies of Aphasic Language. In D.M. Rioch and E.A. Weinstein (Eds.), Disorders of Communication. Baltimore: William and Wilkins, 229-244 (1964).

Porch, B. The Porch Index of Communicative Ability. Palo Alto: Consulting Psychologist Press (1967).

Yorkston, K. and Beukelman, D. A system of quantifying verbal output of high-level aphasics. Proceedings of the Seventh Annual Clinical Aphasiology Conference. Minneapolis: BRK Publishers (1977).

Zurif, E.B. and Caramazza, A. Psycholinguistic structures in aphasia: Studies in syntax and semantics. In W. Whitaker and H.A. Whitaker (Eds.) Studies in Neurolinguistics, Vol. 1. New York: Academic Press, 261-293 (1976).

Discussion

Q: It would be interesting to look at pause time, since normals when they pause seem to hesitate before high information words. Since your data seem to imply that high-level aphasic subjects are having more retrieval problems in grammaticality, it might be interesting to track the reduction of pausing across time as an index of recovery.

A: In developing the rule system for the identification of grammatical strings, we attempted to develop rules which could be applied to all levels of aphasic impairment. Since pause time is probably significant for some people and not for others, we have not included that in our analysis. For example, the anomic speaker whose recovery date I presented would fill in any pause with non-propositional statements. So, despite very obvious word retrieval problems, pause time would not be a particularly useful index of recovery.

Q: Do you have any word fluency measures for speakers?

A: No. At the outset of this project we tried to identify a task where the variability of normal speakers was relatively small. This would allow us to say with some confidence that a low performance is really indicative of the fact that the speaker is functioning outside the normal range rather than just representative of a normal speaker performing poorly on the task. In some pilot work we found that word fluency measures were quite variable even for normal speakers.

Q: How did you handle the fluent speaker whose speech is largely jargon?

A: In this project we were specifically interested in differences that exist between the high level aphasic speaker and normal speakers. With speakers functioning at this level we didn't find a significant amount of jargon. If you examine more severely involved speakers, however, the variability on these measures increase. For example, with moderately severe non-fluent speakers you would expect to find many single word strings and with a moderately severe fluent speaker the string length might be quite long but wouldn't necessarily contain more information.

Q: Were their specific strategies in treatment aimed at the difficulties in grammaticality?

A: No, in fact the treatment given to the three people discussed in this paper was quite different and none received treatment focusing on grammaticality of connected speech.

Q: You seem to be developing a package for analyzing connected speech. Is there some mechanism in the package that deals with truth value, so that if the patient says "the boy is stealing a cookie" the scores are different from someone who says "the boy is chopping a chicken."?

A: Those two sentences are scored similarly on the two measures presented today, because both of the sentences are grammatical. However, they would be scored differently in terms of the amount of the information conveyed, i.e. concepts per minute. This measure has a "truth value" in that the concepts that are used were generated by at least one normal speaker.

Q: If a speaker had said "the boy are stealing a cookie" would that error be reflected in your measures?

A: Yes, there would be a break in grammatical string because there is a grammatical error in that sentence. So the sentence would be scored as a two-word string, [the boy], and a four word string, [are stealing a cookie].

Q: Given the fact that the rules are relatively complex, do you have any measure of inter-judge reliability?

A: Scoring thus far has been done by a committee of three judges who reviewed all of the transcripts and came to consensus on how they should be scored. This allowed us to negotiate the rules for identification of strings. We haven't done any additional inter-judge reliability. However, the most difficult problem seems to be to create a set of rules that are reasonable and that are a true reflection of grammatical adequacy. I feel fairly comfortable that once the rules are established, the inter-judge reliability will be fairly good. We have a measure of intra-judge reliability in which samples were rescored by a judge after an interval of one week. This reliability is fairly high, with a correlation over .95.

Q: Do you have information on how specific speakers vary from picture to picture?

A: To date we have only used the cookie theft picture because we have norms on that particular picture. But variability is the critical issue when applying this technique clinically, especially in light of the fact that samples are potentially very small. It would be very convenient if several pictures elicited the same scores from speakers, but this has yet to be documented. We are in the process of obtaining double samples from a variety of speakers and may find that reliability is compromised when using such small samples. On the other hand, a small sample is much more convenient given the time constraints in a clinical setting.

Acknowledgment

This research was supported in part by RSA Grant Number 16-P-56818.