

Phase I Final Report

NIH Grant # 1 R43 DC009926-01

An Accessible, Effective Treatment for Sentence Deficit in Agrammatic Aphasia

Ronald A. Cole (PI)

March 29, 2010

Rationale for the Proposed Work

The goal of the Phase I research and development effort was to demonstrate the feasibility of developing a commercial prototype of Sentactics, a computer program that could be used independently by individuals with agrammatic aphasia to improve their speech production and comprehension abilities. Sentactics is based on Treatment of Underlying Forms (TUF) a theoretically grounded program that has been demonstrated effective in improving both speech production and comprehension abilities of individuals with agrammatic aphasia. TUF focuses on treatment of syntactically complex sentences, highlighting the movement operations involved in their derivation (as in Government Binding (GB) theory; Chomsky, 1981, 1986). Verbs, which play a central role in grammatical sentences, are carefully selected based on their linguistic properties and entered into target sentences. Using the underlying canonical form of sentences (i.e., subject-verb-object, in English), TUF involves a series of steps focused on the verb and its arguments, or participant roles (e.g., who did what to whom), and how verb arguments move from their canonical position to form the complex structures of interest. Studies examining the effects of TUF, conducted over the past 15 years, show that patients regain ability to comprehend and produce trained sentence structures, and in addition, they show generalization to sentences that are linguistically related, i.e., generalization occurs to sentences that involve similar movement properties. For example, training object cleft structures, such as “It was the artist who the thief chased” results in improved production (and comprehension) of object extracted wh-questions such as “Who did the thief chase?” (Thompson et al., 1997; 1998; 2003).

Development of Sentactics was motivated by the fact that TUF is not currently used widely because the protocols are difficult to administer and require up to 20 sessions to complete. Clinician-administered TUF thus requires considerable training of clinicians and significant costs due to the number of sessions involved. Sentactics was developed to provide an accessible, inexpensive, engaging and effective alternative to clinician administered TUF.

A Tale of Two Companies: How Collaboration with BLT Benefited the Phase I R&D Effort

The SBIR grant was awarded to Mentor Interactive Inc. During the period of the SBIR grant, PI Ron Cole worked half time at Mentor Interactive Inc. and half time at Boulder Language Technologies.

In November 2009 Dr. Cole moved the project staff to Boulder Language Technologies (BLT). The principals at Mentor Interactive and BLT agreed wholeheartedly that moving the R&D effort from Mentor to BLT would accelerate progress towards meeting the Specific Aims by leveraging ongoing software development efforts at BLT, as explained below. The NIDCD program officer, Judith Cooper, was notified of the decision to move the R&D effort to BLT. Although we initially planned to transfer the grant from Mentor to BLT, we decided to simply conclude the research project at BLT without officially transferring the grant (in order to save everyone the effort of moving the grant). Mentor Interactive Inc. continued to administer the grant by providing financial oversight, drawing down and distributing funds to pay project staff, and paying rent to BLT to provide space and infrastructure for the project staff working at BLT.

The main reason for moving the project from Mentor to BLT was to leverage software infrastructure and technologies developed at BLT. Since 2007 BLT has received over \$8 million in research grants and development contracts. Work conducted under these projects has led to the development of a new platform-independent software architecture, the BLT Virtual Human Toolkit BLT VHT), which supports real conversational interaction with a virtual tutor or clinician from PCs or MACs. In addition, BLT has developed new speech recognition and character animation technologies that are needed for Sentactics. By moving the project to BLT, we were able to develop a Sentactics prototype at the end of the Phase I SBIR grant. The commercial prototype runs within the BLT Virtual Human Toolkit system architecture and FlashFace, a new character animation system, described in more detail below. By collaborating with BLT, the project was able to meet and exceed the Phase I Specific Aims, and position BLT for a successful Phase II grant.

Specific Aim 1: Refine the Sentactics Program

The Phase I project had two Specific Aims: To refine the Sentactics program and to investigate the feasibility of using speech and language technologies to provide accurate feedback to clients on the utterances produced during Sentactics applications. The main objective of **Specific Aim I** was to refine the Sentactics program, based on our own observations and interviews of subjects who had used the original version of the program developed at CU. Specifically, in the Phase I proposal we proposed “to modify and refine Sentactics by (a) streamlining specific visual stimuli and prompts, based on the results of our previous research, so the program is easier for patients to use, (b) refining the lexical content and the number of sentence stimuli and corresponding pictures/words used in both the training and test phases of the program, and (c) developing a vocabulary familiarization component that will provide practice with the content words (nouns and verbs) included in the sentence stimuli.”

These objectives were accomplished through collaboration between project staff in Boulder (the PI, software developer and digital artist) and consultants Cynthia Thompson and Audrey Holland. We developed a complete set of new sentences using vocabulary that more clearly differentiates the roles and actions of the characters across sentence types. We also developed, reviewed and selected art work for illustrations that more clearly emphasize the identities and actions of the characters referred to in each sentence type. In addition, we specified the human computer interface for a vocabulary familiarization component that incorporates both comprehension and production of all of the words and phrases used to refer to agent roles actions in the different sentence types used in Sentactics. These refinements are expected to contribute to more satisfactory and engaging user experiences and treatment outcomes for individuals using Sentactics, which we will propose to measure in a Phase II study.

In addition to the work initially proposed under Specific Aim 1, we developed an initial prototype of the Sentactics program based on the BLT Virtual Human Toolkit, a new system architecture developed at BLT that provides a powerful and flexible development environment and runtime platform for future delivery of Sentactics via the internet. As described below, the BLT Virtual Human Toolkit provides an accessible, cost-effective and scalable solution for delivering Sentactics to individuals anywhere and anytime. The initial Sentactics prototype also incorporated FlashFace, a new character animation system developed at BLT that uses motion capture data extracted from video recordings of a human clinician to control, with great accuracy, the facial expressions, head movements and visual speech produced by the virtual clinician. Because of the infrastructure developed at BLT—The BLT Virtual Human Toolkit and FlashFace—it was possible to develop a commercial prototype of Sentactics during two weeks of concentrated effort.

Limitations of the Original Sentactics Program Developed at CU

In Sentactics, a virtual clinician, Sabrina, a lifelike computer character, replaces the human clinician who would administer Treatment of Underlying Forms. During Sentactics activities, Sabrina produces accurate visual speech (movements of the lips, tongue and jaw) synchronized with audio recordings produced by a human clinician (the “voice talent,” a researcher from Dr. Thompson’s lab). The original computer-based Sentactics program was developed by PI Ron Cole and his colleagues at the Center for Spoken Language Research (CSLR) at the University of Colorado (CU) in close collaboration with Cynthia Thompson, the inventor of TUF, and Audrey Holland. The Sentactics system developed at CU was an excellent software program: it was stable; it produced positive user experiences, and published research by Thompson et al., (In Press) demonstrated treatment effects similar to those obtained with human clinicians. The screen shots in Figure 1 show examples of Sabrina in Sentactics sessions.

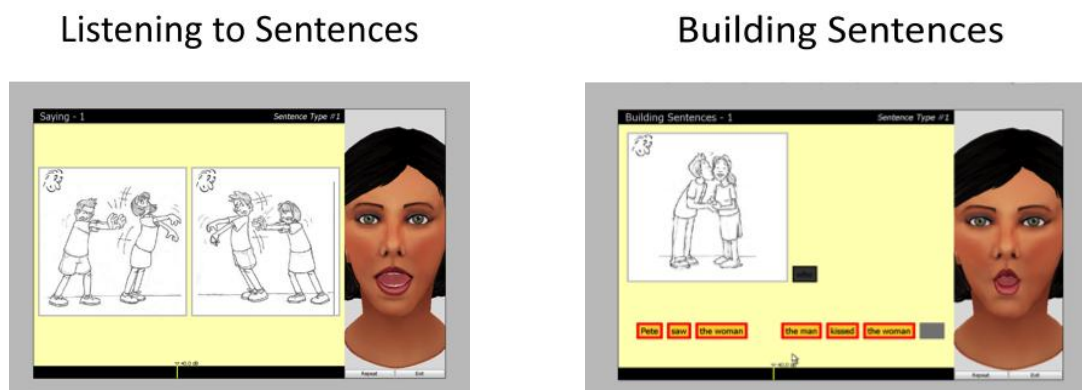


Figure 1: Screen shots of original Sentactics program.

Despite the positive user experiences and treatment outcomes, our analysis of the CU Sentactics system revealed that it could not be extended to become a commercial product. First, the program could only be run on PCs because the character animation system, called CU Animate, which controlled Sabrina, was developed using Microsoft Foundations Classes as an integral part of the software. Thus, Sabrina would only work on machines running Windows. Second, as we learned from hard experience, the CU Animate system was poorly documented, and the original programmer had moved on, so the program could not be modified to produce new 3-D models or to program new facial expressions, head movements or speech behaviors. Finally, the Sentactics program developed at CU worked as a “standalone” program (on PC), whereas there are many benefits of developing a web-based Sentactics program that customers can access from any available PC or MAC.

In order to produce an accessible and scalable program, we decided to develop the Sentactics prototype using the BLT Virtual Human Toolkit (VHT). The BLT VHT system architecture, shown in Figure 2, supports online access to Sentactics from any PC, Mac or handheld device that runs Flash applications. The BLT VHT also supports real time client-server interaction with a lifelike computer character that produces more accurate head movements, facial expressions and visual speech than CU Animate, since all these behaviors are based directly on motion capture data collected from a human speaker. This solution, used in movies such as Shrek and Avatar, produces the most accurate and realistic animation of a computer character. While the cost of animating computer characters using motion capture is expensive (e.g., thousands of dollars per minute in commercial studios), BLT developed its own motion capture solution; this end to end system (from motion capture in our studio to animation of the 3D model based on the mocap data), produces high quality character animation for under \$100 per minute. For an application such as Sentactics, which has less than 100 minutes of speech, the total cost of animating the virtual clinician is approximately \$10,000.

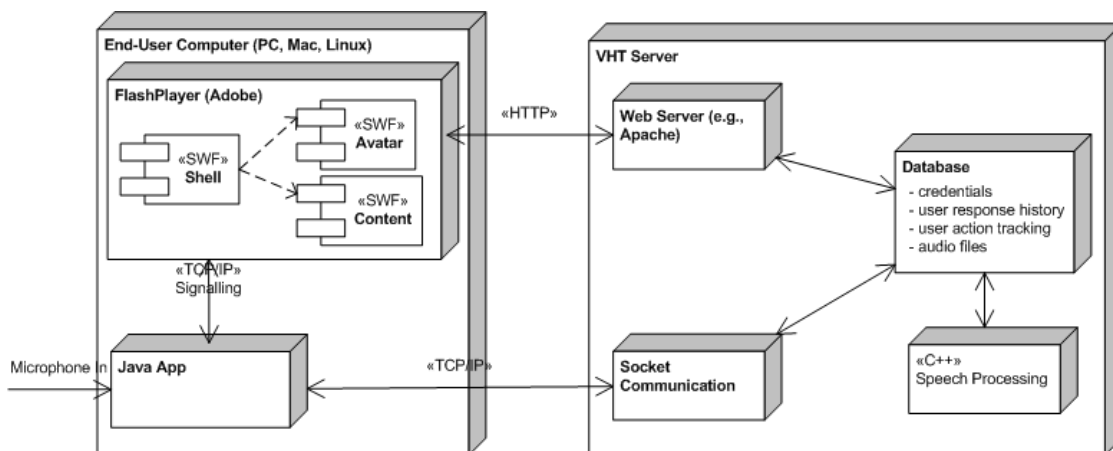


Figure 2: BLT Virtual Human Toolkit Client-Server Architecture.

Using the BLT VHT system architecture and the FlashFace character animation system, we were able to program an initial Sentactics prototype. The system was implemented in March 2010, and successfully tested for a complete Sentactics session on PCs and Macs. Screenshots of Sentactics showing the new avatar, animated using FlashFace, is shown in Figure 3. We plan to put the system online in approximately three weeks, so reviewers of the SBIR Phase II grant proposal can be directed to the BLT Web site to test the prototype if they desire.

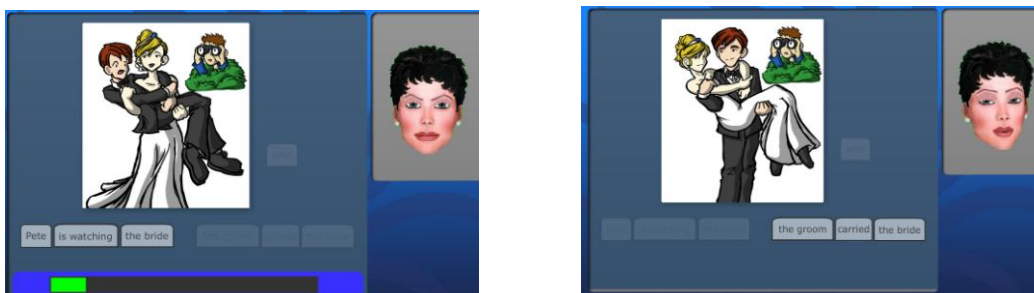


Figure 3: Screen shots of the new Sentactics program with FlashFace avatar.

Specific Aim 2: Develop and evaluate the feasibility of a spoken language system.

The objective of Specific Aim 2, as stated in the Phase I SBIR proposal, was to “investigate the feasibility of using a spoken language system to judge the correctness of utterances produced by individuals while using the Sentactics program. The system was trained and evaluated on utterances produced by individuals with agrammatic aphasia who were recorded while using the beta version of the Sentactics program developed under NIH support.”

Substantial effort was devoted to this aim. At the onset of the project, after weeks of trying, we learned that it was not possible to recover and use the speech that was recorded by subjects during previous tests of Sentactics. We therefore decided to record speech of 6 new subjects. Subjects were recorded by Audrey Holland at the Adler Center in New Jersey. The utterances produced by the six subjects were used in the feasibility study described below.

Rationale for Specific Aim 2

Recent experiments with the Sentactics system developed at CU revealed that the program produced engaging experiences and effective outcomes (Thompson et al., In press). We observed, and subjects consistently reported, that they become immersed in the treatment sessions and that they felt like they were interacting with a real clinician. The virtual clinician is perceived as responsive, effective and supportive. We believe this occurs for several reasons: because the activities are interesting and challenging, the patient is highly motivated, and the virtual therapist is designed to behave like a sensitive and effective clinician. Perhaps the main reason the virtual therapist in Sentactics is so engaging is that she speaks with a natural human voice (recorded by an experienced clinician) rather than with synthetic speech. By using a human voice, we are able to imbue the virtual therapist with a human personality consistent with the full range of expressions and emotions that a human voice can provide.

In a recent study, Thompson et al. (In press) compared treatment outcomes of human administered TUF and Sentactics. The results of these studies revealed equivalent positive outcomes for the two conditions. However, for subjects who used the Sentactics program, a “Wizard of Oz” procedure was used to provide clinician-mediated feedback to the subject on the correctness of their spoken responses. In this procedure, a human clinician monitored each participant’s verbal responses and, using a remote keypad interfaced with the computer, triggered the program to respond with appropriate built-in feedback as to the correctness of a response. Without this feedback, Sentactics operates similarly to other commercially available computer treatment programs in that the patient must use their subjective judgment to determine how closely their verbal responses approximate the computer-presented model.

The objective of research conducted under Specific Aim 2 was to demonstrate the feasibility of incorporating human language technologies within Sentactics (specifically, automatic speech recognition and natural language processing) to provide accurate feedback to individuals on the correctness of their spoken utterances. Successful outcomes of this research would justify additional effort leading to a commercial Sentactics prototype that individuals with aphasia could use independently to improve their comprehension and production of language, consistent with the results of Thomson et al. (2010). The Phase I research study provides strong evidence that human language technologies can be used in Sentactics activities to provide accurate feedback to individuals on the correctness of their utterances.

Subjects

Subjects were 6 adults (5 males, 1 female) with agrammatical aphasia. To preserve confidentiality, subjects are referred to A, B, C, D, E and F throughout the report.

Data Collection Procedure

In the data collection procedure, the prompts were presented by a human experimenter, Audrey Holland. The subject wore a headset microphone. The subjects’ spoken responses were recorded onto a computer. The entire session was recorded as a single audio file. Subjects did not press a button to speak; the recording program continuously recorded all spoken responses. Data were recorded at a sampling rate of 40 KHz using 2 channels.

In the procedure used, subjects produced verbal responses to prompts. The prompts were divided into 10 sections:

1. **Picture Sets** - 4 items. The subject described the action shown in a picture with a sentence ending in a Wh relative clause. The subject is shown 2 pictures and given instructions e.g., *Here are two pictures. For this one you could say “Pete is watching the groom who the bride carried”.* For this one you could say ...
2. **Reading** - 4 items. The subject read sentences ending in Wh relative clauses e.g., *Pete is watching the rabbit who the squirrel kicked.*

3. **Repetition** - 4 items. The subject repeated sentences spoken by the experimenter. Sentences all ended in Wh relative clauses.
4. **Matrix Clause Reading** - 4 items. The subject read sentences of the following form, *Pete is watching the groom*.
5. **Embedded Clause Reading** – 4 items. The subject read sentences of the following form, *The squirrel who kicked the rabbit*.
6. **Matrix Clause Repetition** – 4 items. The subject repeated sentences spoken by the experimenter. Sentences are of the following form, *Pete is watching the groom*.
7. **Embedded Clause Repetition** – 4 items. The subject repeated sentences spoken by the experimenters of the following form, *The squirrel who kicked the rabbit*.
8. **Single Word Production (Action)** – 2 items. The subject was shown a picture and asked to name the action, e.g., *Carry*.
9. **Single Word Production (Agent)** – 4 items. The subject was shown a picture and asked to name the Agent, e.g.: *Who is this?*
10. **Single Word Repetition** – 6 items. The subject was asked to repeat isolated words spoken by the experimenter. The words were the actions and agents from the prior examples.

Experimental Procedure

The goal of automatic processing is to classify each spoken response as Correct or Incorrect relative to a Reference response to the prompt. The experiment measures the Classification Accuracy defined as number of correct responses / total number of responses.

The steps taken in determining the Classifications Accuracy were:

1. Create a Reference Parse for each response representing the Predicate-Argument structure of a correct response using Thematic Roles (See example below).
2. Parse human generated transcripts of spoken subject responses to generate Thematic Roles for the subject response. Correct these manually to guarantee correctness of parse. This is the Transcript Parse.
3. Generate a Transcript Classification for each transcribed response by classifying it Correct if the Transcript Parse matches the Reference Parse, Incorrect otherwise.
4. Transcribe each spoken response using Automatic Speech Recognition (ASR). This is referred to as the Automatic Recognition Hypothesis.
5. Parse Hypothesis for each response into Thematic Roles. This is the Hypothesis Parse.
6. Classify each Hypothesis Parse as Correct if it matches the Reference Parse, Incorrect otherwise. This is the Hypothesis Classification.
7. Compare the Hypothesis Classification for each response to the Transcript Classification. The System Score is True if the Hypothesis Classification agrees with the Transcript Classification, False otherwise.
8. Classification Accuracy is the Number of True System Scores divided by the total number of responses.

Thematic Role Representation

Thematic Roles represent the relations between a Predicate (process) and its Arguments (the entities that are participants in the process). There is not complete agreement among linguistics researchers on a single set of role labels, but for our purposes, the label names don't matter, just our ability to consistently apply the same labels. The responses used in this experiment only need to use 2 roles:

- **Agent** – Usually an agent is a conscious entity that brings about a state of affairs. We do not distinguish between an Agent and an Experiencer, but combine both under the Agent label.
- **Theme** – The entity that is affected by the content of an experience.

Since there are very few predicates used in this task, it was useful to create more specific roles for the predicate *Watch*; i.e., *Watcher* and *Watchee*. Normally the predicate is just given the label *Predicate*, but since the number of predicates is small and they are referred to as Actions in the protocol, the predicate is given the label *Action* (for predicates other than *Watch*). For example: *Pete is watching the groom who the bride carried* would be represented as

Watch: watch **Watcher:** Pete **Watchee:** groom
Action: carried **Agent:** bride **Theme:** who

Traditionally, the relative clause would be included in the Theme of *watching* (i.e., Theme: the groom who the bride carried). But for our purposes, scoring is greatly simplified without affecting the results by not including the relative clause or the determiner *the*, but only the entity groom as the Theme of watching. In the context of scoring the correctness of utterances produced by subjects in the Sentactics project, the objective is to determine whether the right entity is in the right role relative to each predicate.

The Phoenix semantic parser was used to generate the Thematic Role parses. Phoenix is a rule based system that uses semantic grammars to map word strings directly onto semantic frames (Ward 94). Since the entities and predicates involved in our experiment were each a fixed set, a small set of rules was sufficient to provide complete coverage of the strings to be parsed. Phoenix does not parse every word in a string, just the content phrases, so it is very useful for parsing input that contains disfluencies, insertions and repeated words and phrases.

The set of entities and predicates used in the experiment is:

Entities: Pete, bride, groom, squirrel, rabbit.

Predicates: watch, carry, kick.

The process of generating Reference Classifications is shown in Figures 1 and 2.

Figure 4 shows an example where the System Classification is Correct (the system would give the same feedback as a human experimenter):

- The Reference Answer (the correct answer) for the prompt is *Pete is watching the rabbit who the squirrel kicked.*
- The Reference Parse puts the Entities *pete* and *rabbit* (and the anaphoric reference *who*) into their roles relative to the predicates *watching* and *kicked*.
- The Transcript (what the subject actually said) was, *uh Pete is watching who uh who the rabbit who the squirrel kicked.*
- The Transcript Parse puts the entities in roles relative to the predicates.
- The Transcript Parse is compared to the Reference Parse and is found to be the same, so the transcript is classified as Correct.
- The recognition Hypothesis (output of the speech recognizer) for the subject speech was *for i a pete is watching who a the a who a the rabbit who the the who a the squirrel kicked.*
- The speech recognition Hypothesis is parsed into roles (the Hypothesis Parse).
- The Hypothesis Parse is compared to the Reference Parse and is found to be the same, so the Hypothesis is classified as Correct.
- The Hypothesis Classification is compared to the Transcript Classification, and the two agree so the System Score is True. Note: this does not mean that the system said the subject was correct. It means that the system would give the same feedback to the user that the human experimenter would give, so the system is judged as correct in the response that it generated.

Reference Answer: <i>pete is watching the rabbit who the squirrel kicked</i>
Reference Parse:
[Watcher].pete
[Watch].watch
[Watchee].rabbit
[Agent].squirrel
[Action].kick
[Theme].who
Transcript: <i>uh pete is watching who uh who the rabbit who the squirrel kicked</i>
Transcript Parse:
[Watcher].pete
[Watch].watch

[Watchee].rabbit

[Agent].squirrel

[Action].kick

[Theme].who

Transcript Class: C

Hypothesis: *for i a pete is watching who a the a who a the rabbit who the the who a the squirrel kicked*

Hypothesis Parse:

[Watcher].pete

[Watch].watch

[Watchee].rabbit

[Agent].squirrel

[Action].kick

[Theme].who

Hypothesis Class: C

System Score: True

Figure 4: Example of a correct System Score by the system.

Figure 5 shows an example where the system makes a mistake and would have given wrong feedback to the subject:

- The Reference Answer for the prompt is *the bride carried the groom*.
- The Reference Parse puts the Entities *bride* and *groom* into their roles relative to the predicate *carried*.
- The Transcript was *uhm bride carries groom*.
- The Transcript Parse puts the transcript entities in roles relative to the predicate.
- The Transcript Parse is compared to the Reference Parse and is found to be the same, so the transcript is classified as Correct.
- The recognition Hypothesis was *the a it carries it groom*.
- The speech recognition Hypothesis is parsed into roles (the Hypothesis Parse).
- The Hypothesis Parse is compared to the Reference Parse and is found to be missing the element *[Agent].bride*, so the Hypothesis is classified as Incorrect.
- The Hypothesis Classification is compared to the Transcript Classification, and since the two disagree the System Score is False the system would give incorrect feedback to the subject.

Reference Answer: *the bride carried the groom*

Reference Parse:

[Agent].bride

[Action].carry

[Theme].groom

Transcript: *uhm bride carries groom*

Transcript Parse:

[Agent].bride

[Action].carry

[Theme].groom

Transcript Class: C

Hypothesis: *the a it carries it groom*

Hypothesis Parse:

[Action].carry [Theme].groom Hypothesis Class: I System Score: False

Figure 5: Example of an Incorrect Classification by the system.

Figure 5 illustrates the most common type of error made by the system. The speech recognition failed to correctly recognize one entity which caused a single role to be missing from the parse.

Corpus Development

The recorded speech from the six subjects was processed to create a corpus:

- Audio files for each session were converted from 44.1kHz, 2 channel .aiff files into 16kHz, 1 chan .raw files, which is the format required by our ASR system.
- Each session was recorded as one long file with experimenter speech and subject speech intermixed. The file was segmented into turns, each representing a subject response to a prompt presented by the experimenter. In order to produce an accurate estimate of how well the automatic system can assess the correctness of the subjects' spoken utterances in a Sentactics application, we needed to segment each subject's speech file into turns and then process the subject's speech one turn at a time. The audio files were hand transcribed to get reference word strings (.trans files). The transcripts contained only subject speech (not experimenter speech). Each turn of subject speech contained start and end time stamps which were used to segment the audio files into turns. The transcripts were also segmented into turns corresponding to each audio file. Disfluencies were also marked in the transcripts. These were not used in speech recognition, but were only used to inform error analysis. The disfluency annotations were filtered out before training language models or calculating the word error rate. In other words, the natural language processing system filtered out the disfluencies, as would a human listener, in order to determine if the words produced by the subject contained the correct thematic roles and relationships.
- Transcripts of each individual turn were aligned with the reference responses from the protocol, that is, each transcript was annotated to indicate which protocol item (1-40) the utterance was responding to.

Speech Processing

The Sonic speech recognition system (Pellom 2001; Pellom & Hacioglu 2003), developed at the University of Colorado was used to generate the word hypotheses for utterances produced by each subject. Sonic matches phone level acoustic models used by the system against the speech input from the user. It combines the results of the acoustic match with language models which provide a statistical representation of the expected co-occurrence of words. The combined acoustic and language model information is used to select the most likely sequence of words corresponding to the speech input.

The acoustic models used in the automatic system were our standard models, which have been trained on a large corpus of male and female adult speakers. The speech produced by the six subjects in our experiment was not used to train the acoustic models, and no speakers with aphasia were in the set of speakers used to train the system. No specific provision was made in the speech recognition system to deal with disfluencies. As mentioned earlier, the Phoenix parser is particularly robust in processing disfluent and errorful input.

A Trigram Language Model was trained for the application. That is, statistics about the expected sequence of words was based on sequences of word triplets as described below. For training the Trigram Language Model speakers were divided into a training set and a test set.

- Training set: A, E
- Test set: B, D, F, C

The LM training data were created from transcripts of training speakers and reference answers to protocol items. To make the model more robust, all verb inflections were mapped to a base form (ie, carry, carries, carried and carrying were all mapped to carry). These were all treated as alternate pronunciations of the baseform in the recognizer lexicon. The recognizer lexicon was 84 words.

Scoring Speech Recognition Accuracy

The standard measure of automatic speech recognition performance is Word Error Rate, which is the sum of Insertion, Substitution and Deletion errors. For our purposes, Insertions are generally harmless, and due to the large numbers of disfluencies and some utterances by some subjects there are often many insertions. To include them would not accurately represent the content of the recognition output for use in this task. We therefore measured the Word Accuracy of the speech recognition output. Word Accuracy is calculated by aligning speech recognizer output with the reference transcript and is the percentage of words in the transcript that were correctly recognized. The word accuracy score was computed using standard recognition scoring software (the NIST SCLITE routine).

Results

Table 1. – Overall Classification Accuracy by Subject.

Subject	Word Accuracy	Subject Errors	Classification Errors	Items Done	Classification Accuracy
A	0.73	39	1	40	0.97
B	0.90	5	2	36	0.94
C	0.78	11	5	38	0.86
D	0.67	18	7	40	0.82
E	0.63	10	7	40	0.82
F	0.54	18	8	40	0.80

Table 2. – Classification Accuracy for Each Section of Protocol.

Subject	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
B	1.0	1.0	-	1.0	0.75	0.75	1.0	1.0	1.0	1.0
D	1.0	1.0	1.0	0.75	0.75	0.75	0.75	0.50	0.75	0.83
F	1.0	1.0	1.0	0.75	0.50	0.75	0.25	1.0	1.0	0.83
C	1.0	0.50	1.0	0.50	0.75	1.0	1.0	0.50	1.0	1.0
A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.50	1.0	1.0
E	1.0	0.75	1.0	1.0	1.0	0.75	0.50	1.0	0.50	0.83
ALL	1.0	0.91	1.0	0.83	0.79	0.83	0.75	0.75	0.88	0.92

Table 3. – Subject Response Accuracy for Each Section of Protocol.

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
0.04	0.14	0.10	0.67	0.63	0.79	0.79	0.83	0.75	0.71

Table 4. – Word Accuracy for Each Section of Protocol

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
0.59	0.69	0.66	0.81	0.82	0.86	0.76	0.70	0.77	0.89

Table 1 shows the overall System Classification Accuracy by subject along with Word Accuracy for that subject. While there is a general trend that Classification Accuracy is higher for subjects where the recognition Word Accuracy was higher, the one exception to the trend was the subject A, where the system had the highest Classification Accuracy. The Subject Errors column is the number of items that the subject got wrong. All System Classification errors were False Rejections, i.e., the subject gave a correct response, but the system classified the response as Incorrect. This was due to the speech recognizer not recognizing a content word. There were no cases in which the subject gave an incorrect response, but the system classified it as Correct. This explains the high System Classification accuracy for subject A. Since subject A missed 39 of the 40 items, and the system made no False Acceptance errors, it had a high classification accuracy.

Subject B made only a few errors earning a high word accuracy, also yielding a high System Classification Accuracy. This illustrates a general trend in the data. As subjects are having a more difficult time with the task, they are more disfluent in their speech. The more disfluent speech gives a lower Word Accuracy from the recognizer, but the subject is also more likely to respond incorrectly. Since the system made no False Acceptance errors (speech recognition errors did not cause the system to classify an incorrect response as correct) the system maintains relatively good classification accuracy in the presence of low word accuracy from the speech recognizer.

This positive correlation between subjects' disfluencies and the difficulty they are having with the task suggests that speech recognition and natural language processing can accurately classify subjects' responses in the Sentactics task as correct or incorrect. Word accuracy and System Classification are good when subject's speech is relatively fluent. When subjects' speech is very disfluent and problematic for the recognition system, it is also the case that the subject is having great difficulty with the task and is very unlikely to produce a correct response. As long as the system does not make False Accept errors, this situation also gives high system classification accuracy. In proposed Phase II research, we will test this hypothesis with a much larger pool of subjects.

Tables 2, 3 and 4 reinforce this point. Table 2 shows System Classification Accuracy for each set of protocol items. Table 3 shows the percentage of subject responses that were correct for each of these sets. Table 4 shows recognizer Word Accuracy for each set. Sets 1, 2 and 3 were the hardest for the subjects in that they had lowest subject response accuracy. These sets also have low word accuracy. However, all three sets have high system classification accuracy. Sets 9 and 10 have very short, often single word answers, and therefore contain few disfluencies. This is a much easier task for the subjects and the speech recognizer, and the word accuracy is relatively high. Subject responses in these sets are mostly correct, but the word accuracy is high so the system is still able to classify the response correctly.

Discussion

The trend shown in the data-- that the degree of disfluencies in utterances is correlated with incorrectly produced responses in terms of thematic role assignments-- allows for reasonable system classification rates across subjects and types of prompts. The places that the speech recognition made mistakes are mostly responses that subjects got wrong.

We are confident that word accuracy will be improved significantly by the application of techniques to adapt the speech recognition system to the idiosyncratic features of this task. Proven techniques that we hope to implement in Phase II research can be applied to reduce errors. These include:

- **Disfluencies** – There are many disfluencies in the data: restarts, truncated words, clearing throat, laughing, loud sighs, etc. Word accuracy could be improved by explicit modeling of these disfluencies with so called “garbage models”. Training these models requires more data than we had available. With sufficient data, such models can be trained and can significantly improve the performance of the speech recognizer.
- **Task Specific Acoustic Training** – The acoustic models used by speech recognizers model phonemes in the context of the preceding and following phonemes. In principle, these models are general and not specific to any lexicon or task. In practice, gathering data from the target task and adapting the models to the task specific speech improves performance. This is especially important in cases where users have speech characteristics not present in the data the models were trained on; such as accents, or in this case, speech from aphasic subjects. The small amount of data available for this experiment was not sufficient for acoustic training. The collection of a corpus from subjects engaged in the Sentactics task would allow acoustic model adaptation to the task.
- **Pronunciation** – Much of the speech used for this experiment contains reasonable pronunciations, but there are still quite a few pronunciation errors that contribute to the error rate (for example, *groom* for *broom* and *curl* for *squirrel*). If an adaptation procedure were to be added where alternate pronunciations were added to the lexicon and acoustic models were adapted to allow for more reduction and variation, word accuracy could be improved. In large vocabulary speech recognition tasks, this is usually not an effective strategy because it decreases the ability of the system to distinguish between acoustically similar words. In very small vocabularies like the one used in the Sentactics task, pronunciation constraints can be reduced as long as it doesn't make the word confusable with another word in the lexicon that would tend to be used in the same context (or with a garbage model if they are used). In this study, words in the transcripts that were close to the intended word, were mapped to the intended word and treated as correct in the transcript. For example, *broom* was mapped to *groom*, *gride* was mapped to *bride* and *curl* mapped to *squirrel*. So those words are treated as correct in the transcript processing. However, the speech lexicon did not contain *broom*, *gride*, or *curl* and the recognizer either had to produce the correct word (even though it wasn't pronounced correctly) or are scored as incorrect. Speech recognition systems are sensitive to this type of pronunciation error, but in small lexicons with non-confusable words, the pronunciation models can be broadened.

- **Dynamic range** – In some recordings, the dynamic range of the recording was very large. The sentence might start off with very soft speech and then get much louder. The very soft speech was often treated as background sounds (rather than speech produced by the subject) by the normalization routines used by the recognizer and caused the word to be missed. Simply boosting the volume of the recording isn't a solution, because the loud parts would then clip. As with acoustic phone models, the background normalization and Voice Activity Detection models can be adapted with sufficient training data. Another simple solution is to have the system give the subject feedback when they are speaking too softly or too loudly. A sound pressure meter can be added to the display that shows the subject when their volume is in an acceptable range. This technique is often used in spoken language interfaces to good effect.

References

- Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin & use*. New York: Praeger.
- Pellom, B. (2001) "SONIC: The University of Colorado Continuous Speech Recognizer", University of Colorado, tech report #TR-CSLR-2001-01, Boulder, Colorado, March.
- Pellom, B., Hacıoglu, K. (2003) "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, April.
- Thompson, C., Shapiro, L., Ballard, K., Jacobs, B., Schneider, S., & Tait, M. (1997). Training and generalized production of wh-and NP-movement structures in agrammatic aphasia. *Journal of Speech, Language, and Hearing Research*, 40, 228-244.
- Thompson, C. K., Ballard, K., & Shapiro, L. (1998). The role of syntactic complexity in training wh-movement structures in agrammatic aphasia: Optimal order for promoting generalization. *Journal of the International Neuropsychological Society*, 4, 661-674.
- Thompson, C. K., Shapiro, L., Kiran, S., & Sobecks, J. (2003). The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia: The complexity account of treatment efficacy (CATE). *Journal of Speech, Language, and Hearing Research*, 42, 690-707.
- Cynthia K. Thompson, JungWon Janet Choy Audrey Holland & Ronald Cole (In press) Sentactics@: Computer-Automated Treatment of Underlying Forms. *Journal of Aphasiology*.
- Ward, W., "Extracting Information From Spontaneous Speech", *Proc. ICSLP-94*, September 1994.