AphasiaBank: Preliminary Lexical, Morphosyntactic, and Error Analyses

BACKGROUND

AphasiaBank collects and analyzes samples of the discourse of individuals with aphasia and normal participants across a range of tasks. One goal of this work is to improve the treatment of aphasia. To reach that goal, we must solidify the empirical database supporting our understanding of communication in aphasia. Toward that end, we have organized a consortium of 80 members who are using a shared methodological and conceptual framework for the processes of recording, transcribing, coding, analysis, and commentary. Our eight specific aims are:

1. Protocol standardization. We have developed a standardized data collection protocol for all consortium sites.
2. Database development. Transcriptions use the CHAT standard in which the transcripts are linked to digitized audio and video.
3. Analysis customization. We have constructed tools for the analysis of multimedia transcripts on the levels of phonology, lexicon, morphology, syntax, discourse, and pragmatics.
4. Measure development. We use the annotations produced by these tools to automatically compute the measures described in this paper.
5. Syndrome classification. Using these new measures and the growing database, we will develop new approaches to syndrome-based patient classification and diagnosis.

6. Qualitative analysis. We are conducting qualitative analysis on a subset of the transcripts with an emphasis on gesture coding and interaction.
7. Profiles of the recovery process. Using the measures we describe, we will trace changes across time in both individual patients and patient groups.
8. Evaluation of treatment effects. We intend ultimately to evaluate the effectiveness of specific aphasia treatments.

METHOD

This paper will illustrate the analyses possible with AphasiaBank data by providing comparative data on a subset of the AphasiaBank collection to date. Participants include 15 normal adults and 15 individuals with aphasia secondary to stroke. Extensive demographic data were collected on all participants. Normal adults were screened for depression, dementia, and history of neurologic and/or communication disorders. All participants had auditory and visual abilities adequate for testing and spoke English as their first language. Basic demographic data are given in Table 1.

Table 1. Demographic Data on Participants

	Normal Adults	**Individuals with Aphasia**
Age range (yrs)	23-79 (mean = 60)	35-78 (mean=64)
Handedness	right =13, left=1 ambidextrous=1	right=13, left=1 ambidextrous=1
Education range (yrs)	12-20 (mean=16)	12-22 (mean=14)

Data were collected using the AphasiaBank protocol that was described at CAC last spring. The protocol was finalized in Spring 2008 and is available from http://talkbank.org/AphasiaBank. The protocol collects data from these four activities:

1. Free Speech Sample. This sample is based on a recounting of the stroke or a similar medical event (for normal adults) and an important life event (Labov & Waletzky, 1967).
2. Picture descriptions of the broken window sequence, the refused umbrella sequence, a complex cat rescue picture (Nicholas & Brookshire, 1993, 1995), and an emotive flood rescue picture.
3. Narrative. The participant is asked to retell the Cinderella story (Saffran, Berndt, & Schwartz, 1989; Webster, Franklin, & Howard, 2007).
4. Procedural discourse. The participant describes how to make a peanut butter and jelly sandwich.

In addition, the protocol includes scores on the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 2001), Northwestern Verb Naming Test (Thompson, in preparation), and Western Aphasia Battery (Kertész, 2007), and a repetition test based on Martin and Gupta (2004) and Holland (unpublished data, 2000).

In this paper, we analyze the four activity components, focusing on lexical, morphosyntactic, and error analyses. For analysis, we use tools created by the TalkBank project (http://talkbank.org), as shaped for the specific goals of AphasiaBank. A team of experienced aphasia researchers is responsible for transcription and coding.

Lexical Analyses. Our goal here is to quantify the participant's control of lexical diversity. Raw lexical diversity is measured automatically by running the VOCD program on the four discourse segments (free speech, picture descriptions, narrative, and procedural discourse). The VOCD measure (Malvern, Richards, Chipere, & Purán, 2004) provides a statistically balanced replacement for the earlier TTR measure that was excessively skewed by sample size. In addition to the VOCD analysis, we have constructed a novel form of automatic analysis based on naming of specific lexical targets, called TLEX. To compute this analysis, we specify a target set of lexical forms from the domains sampled by the four oral components of the protocol. For example, for the Cinderella narrative, we look for use of forms such as “prince”, “stepsister”, “slipper” and so on. Using this target lexical set, we derive scores measuring the extent to which each oral segment matches the target lexicon.

Morphosyntactic Analyses. The TalkBank system provides a complete set of tools for automatic morphosyntactic analysis of language samples transcribed according to CHAT guidelines (MacWhinney, 2008). The first step in the analysis involves part of speech tagging using the MOR program. Next, the analyst runs the POST program to automatically disambiguate any ambiguous forms. At this level, the analysis of aphasic language production can be very close to that for normal language production, since the error coding system (see below) includes methods for inserting target forms into the transcript where appropriate. After running MOR and POST, the transcript has a complete %mor line that tags each word for its morphological form. Next, the analyst runs the GRASP program (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2007) to automatically link together words in terms of grammatical relations. At this point, methods for analyzing aphasic language diverge from those for normal speech. We will discuss how to implement methods to deal with these divergences. For each subject, we present grammatical relations profiles computed by our automated versions of IPSyn (Scarborough, 1990) and DSS (Lee, 1969).

Error Analyses. The AphasiaBank error analysis system, available from http://talkbank.org/AphasiaBank/errors.doc, is a simplification of the system published in MacWhinney (1990). The newer system is more closely tailored to the actual distribution of error types occurring in the transcripts from aphasic individuals we have coded to date.

Scores Being Presented. The tables we will present in the Powerpoint for this talk are based upon these twelve automatically computed numerical measures for each participant:

1. VOCD: lexical diversity,
2. TLEX: percentage hits to target lexicon,
3. MLUwords: mean utterance length of the participant's productions in words,
4. MLUmorphemes: mean utterance length of the participant's productions in morphemes,
5. DSS: score on the DSS (Developmental Sentence Score),
6. SemError: semantic paraphasias, neologisms, jargon, empty speech, and circumlocution,
7. PhonError: phonemic paraphasias (real words and non-words), and
8. GramError: errors in agreement, part of speech selection, missing words, and overall agrammatism.

In addition to these eight corpus-based scores, we will present these four repetition scores.

1. Word Repetition 1 – lists of increasing length from a closed set of words presented in identical order; the score is the longest list correctly repeated.

2. Word Repetition 2 – lists of increasing length from an open set of words; the score is the longest list correctly repeated.
3. Sentence Repetition 1 – sentences increasing in length by adding words to the same core sentence; the score is the longest sentence correctly repeated.
4. Sentence Repetition 2 – sentences with no errors, semantic errors, or an interference effect; the score is the total words correct overall.

These 12 scores, along with the 45 demographic variables we are tracking, are the input to a general cluster analysis designed to generate a new data-based classification of aphasia types. However, for this presentation, our graphs will only display these variables across aphasic types provided from the clinical interviews.

References:

Adamson, L., Bakeman, R., & Deckner, D. (2004). The development of symbol-infused joint engagement. *Child Development*, 75, 1171-1187.

Bakeman, R., & Adamson, L. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, 55, 1278-1289.

Frederiksen, J. (1981). Understanding anaphora: Rules used by readers in assigning pronominal referents. *Discourse Processes*, 4, 323-347.

Frederiksen, J., Sipusic, M., Sherin, M., & E., W. (1998). Video portfolio assessment: Creating a framework for viewing the functions of teaching. *Educational Assessment*, 5(4), 225-297.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test*. Austin, TX: Pro-Ed.

Kertész, A. (2007). *Western aphasia battery*. San Antonio: PsychCorp.

Labov, W., & Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. In J. Helm (Ed.), *Essays on the verbal and visual arts* (pp. 12-44). Seattle, WA: University of Washington Press.

Lee, L. (1969). *Northwestern Syntax Screening List*. Evanston: IL: Northwestern University Press.

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (pp. 165-198). Amsterdam: John Benjamins.

MacWhinney, B., & Snow, C. (1990). The Child Language Data Exchange System. *ICAME Journal*, 14, 3-25.

Malvern, D. D., Richards, B. J., Chipere, N., & Purán, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.

Martin, N., & Gupta, P. (2004). Exploring the relationship between word processing and verbal short-term memory: Evidence from associations and dissociations. *Cognitive Neuropsychology*, 21, 213-228.

Nicholas, L., & Brookshire, R. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research*, 36, 338-350.

Nicholas, L., & Brookshire, R. (1995). Presence, completeness and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research*, 38, 145-156.

Saffran, E., Berndt, R., & Schwartz, M. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, 37, 440-479.

Sagae, K., Davis, E., Lavie, E., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics*. Prague: ACL.

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

Webster, J., Franklin, S., & Howard, D. (2007). An analysis of thematic and phrasal structure in people with aphasia: What more can we learn from the story of Cinderella? *Journal of Neurolinguistics*, 20, 363-394.