Aphasiology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/paph20

AphasiaBank: Methods for studying discourse
Brian MacWhinney a, Davida Fromm a, Margaret Forbes a & Audrey Holland b
a Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA
b Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, AZ, USA
Available online: 22 Sep 2011

To cite this article: Brian MacWhinney, Davida Fromm, Margaret Forbes & Audrey Holland (2011): AphasiaBank: Methods for studying discourse, Aphasiology, 25:11, 1286-1307

To link to this article: http://dx.doi.org/10.1080/02687038.2011.589893

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
AphasiaBank: Methods for studying discourse

Brian MacWhinney¹, Davida Fromm¹, Margaret Forbes¹, and Audrey Holland²

1Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA
2Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, AZ, USA

Background: AphasiaBank is a computerised database of interviews between persons with aphasia (PWAs) and clinicians. By February 2011 the database had grown to include 145 PWAs and 126 controls from 12 sites across the United States. The data and related analysis programs are available free over the web.

Aims: The overall goal of AphasiaBank is the construction of a system for accumulating and sharing data on language usage by PWAs. To achieve this goal we have developed a standard elicitation protocol, and systematic automatic and manual methods for transcription, coding, and analysis.

Methods & Procedures: We present sample analyses of transcripts from the retelling of the Cinderella story. These analyses illustrate the application of our methods for the study of phonological, lexical, semantic, morphological, syntactic, temporal, prosodic, gestural, and discourse features.

Outcomes & Results: AphasiaBank will allow researchers access to a large, shared database that can facilitate hypothesis testing and increase methodological replicability, precision, and transparency.

Conclusions: AphasiaBank will provide researchers with an important new tool in the study of aphasia.

Keywords: Aphasia classification; Transcript analysis; Corpus analysis; Elicitation protocols; Gesture analysis; Databases.
as well as 126 control or comparison participants. The access to transcripts and video materials is password-restricted to AphasiaBank members, but membership is automatically granted to all researchers studying aphasia on request. Access to the programs, manuals, and other resources is open and free to everyone.

In this article we explain how to use the CLAN (Computerised Language Analysis) programs (MacWhinney, 2000) to analyse phonological, lexical, phonological, morphosyntactic, discourse, and gestural patterns in the database. Several publications have already made use of the AphasiaBank database and CLAN programs. MacWhinney, Fromm, Holland, Forbes, and Wright (2010) conducted various lexical analyses of the segment of the protocol in which PWAs describe the Cinderella story. Fergadiotis, Wright, and Capilouto (2011) examined lexical diversity in younger versus older participants across different discourse types using CLAN analyses, but not with AphasiaBank data. Fergadiotis and Wright (2011) used similar methods to analyse AphasiaBank protocol data. Finally, Fromm et al. (2011) studied responses of PWAs to queries about their speech. Segments of these publications will be used as illustrations of the types of analyses that can be conducted.

FORMATION OF THE APHASIA BANK SYSTEM

AphasiaBank has been designed to replicate and extend the organisational model established by the Child Language Data Exchange System (CHILDES) for the field of child language acquisition. The CHILDES Project, directed by Brian MacWhinney and funded by NIH/NICHD since 1987, is an international cooperative venture involving over 800 active users and 4000 affiliated members located in over 30 countries. Most new empirical studies of child language production rely on the analysis of data from the CHILDES database and the majority of theoretical papers on language that make reference to production data are now based on the use of the CHILDES database. A recent count located over 3500 published articles based on the use of CHILDES data or programs. The system provides users access to a set of programs (CLAN), a database (CHILDES), a transcription system (CHAT), documentation, and an electronic discussion group (chibolts@googlegroups.com) for communicating on issues in language analysis. The form of these tools has been shaped by continual input from active members of the system.

Work on establishing a system of this type for the study of aphasia started in 2005 with a planning meeting of 20 senior aphasia researchers. At this meeting participants specified the shape of the AphasiaBank protocol and outlined methods for data sharing and possible computational analyses. The AphasiaBank grant, prepared by Brian MacWhinney and Audrey Holland, was funded in 2007. Workshops with senior aphasia researchers have continued on a yearly basis to advise on a large number of issues including protocol development, language transcription, error coding, discourse analysis strategies, and future directions. The AphasiaBank website at http://talkbank.org/AphasiaBank/ is the primary source for all AphasiaBank related materials (e.g., transcripts, videos, computer programs, manuals, transcription Center), Gretchen Szabo (Adler Aphasia Center), Roberta Elman (Aphasia Center of California), Denise McCall (SCALE), and Maura Silverman (TAP) for helping us administer the AphasiaBank protocol to the participants at their respective institutions.

Thanks to Heather Wright, Gilson Capilouto (University of Kentucky), and Dan Kempler for their contributions of protocols for control participants.
training, IRB guidelines, ground rules). In addition an AphasiaBank Google Group (currently with 98 members) is used for purposes of information dissemination and discussion of topics relevant to the project.

**THE APHASIA BANK PROTOCOL**

The central goal of the AphasiaBank project is the creation of a shared database of multimedia interactions for studying communication in aphasia. Because of the diversity of clinical patterns in aphasia we considered it important to implement a standard protocol to achieve maximal comparability across participants. To that end we have developed a tightly specified data collection protocol that is being consistently implemented at all participating AphasiaBank research sites. The protocol consists of four different discourse elicitation tasks: personal narratives, picture descriptions, story telling, and procedural discourse. We chose to focus on narrative and procedural discourse in order to maximise task comparability across participants. In the future we expect to include additional methods for collecting conversational discourse.

A script was developed to keep the prompts consistent across investigators. The script includes a second-level prompt to use if a participant does not respond in 10 seconds. A troubleshooting script is also available for participants who still cannot respond and need additional prompting with simplified questions. To maximise comparability across sessions the investigator makes every effort to be as silent as possible during administration of the protocol, while giving maximal non-verbal encouragement. Participants are given as much time as they need for their responses. The protocol is administered in a single session and the session is recorded on video, using a set of guidelines to maintain high audio and video recording quality. These guidelines, which are posted at the AphasiaBank website, specify details regarding the video equipment to be used, the configuration of the equipment, and methods for creation of computer files from the video output. There are four discourse tasks:

1. **Personal narratives.** These are elicited by asking the PWAs about their speech, their stroke, their recovery, and an important event in their lives. Control participants are asked about an illness or injury, their recovery from that illness or injury, any experience they have had with people who have trouble communicating, and an important event in their lives.
2. **Picture descriptions.** Participants are shown three black and white drawings. They are asked to look at the picture and tell a story with a beginning, middle, and end (Wright & Capilouto, 2009). The first picture stimulus is a four-panelled picture of a child playing with a soccer ball and breaking a window, the second is a six-panelled picture of a child refusing an umbrella and getting caught in the rain, and the third is the Nicholas and Brookshire (1993) picture of a cat stuck in a tree. A fourth picture, a colour photo of a flood rescue scene, was used for the first 2 years of the project and then discontinued because many participants were having trouble interpreting the picture.
3. **Story telling.** Participants are shown a paperback picture book of Cinderella (Grimes, 2005), with the words covered up. They are told to look through the book to remember how the story goes. Then the book is taken away and they are asked to tell as much of the story as they can.
4. **Procedural discourse.** Participants are asked to describe how they would make a peanut butter and jam [jelly] sandwich. (Test sites outside the United States may
substitute another simple food preparation.) A stimulus picture with photographs of peanut butter, bread, and jam [jelly] is available for use with participants who need extra help.

Although the current samples are all in English, samples are also being collected in Cantonese, Mandarin, German, and Swedish. Inclusion criteria for PWAs have been limited (with few exceptions) to individuals whose aphasia results from a stroke that can be verified through neuroimaging or a clear medical diagnosis. Extensive demographic data (51 fields) have been collected on each participant and are available to AphasiaBank members at the website. A total of 10 of the PWAs had a second administration of the protocol done approximately 1 year following the initial administration.

In addition to the demographic data, three standardised measures are administered to PWAs: (1) the Aphasia Quotient (AQ) subtests from the Western Aphasia Battery-Revised (WAB; Kertész, 2007); (2) the short form of the Boston Naming Test-Second Edition (Kaplan, Goodglass, & Weintraub, 2001); and (3) the Verb Naming Test from the Northwestern Assessment of Verbs and Sentences-Revised (Thompson, 2011). We also administer a non-standardised repetition test, developed to assess word-level and sentence-level repetition skills. All testing, with the exception of the WAB, is recorded on video. The control participants are tested with the Mini-Mental State Exam (Folstein, Folstein, & Fanjiang, 2002) and the Geriatric Depression Scale (Brink et al., 1982) to rule out cognitive impairment and depression. Test results are also available to AphasiaBank members in a master spreadsheet on the AphasiaBank website.

Transcription and coding

Before being included in AphasiaBank the discourse samples from both PWAs and controls go through a detailed process of transcription, coding, and checking. Transcription is done using the CHAT format (MacWhinney, 2000) that has been developed over the last 30 years for use in a variety of disciplines such as first language acquisition, second language acquisition, classroom discourse, and conversation analysis. CHAT is designed to operate closely with the CLAN programs, which are also described in MacWhinney (2000). These programs allow for the analysis of a wide range of linguistic and discourse structures, some of which will be described in this article. The CLAN program, along with updated electronic versions of the CHAT and CLAN manuals, can be downloaded from the AphasiaBank website at http://talkbank.org/AphasiaBank. That site also provides a transcription training manual that was prepared specifically for AphasiaBank purposes.

For detailed transcription we rely on CLAN’s Walker Controller function that allows the transcriber to replay segments of the video. This replaying can be done with a variable window and a variable lag. When desired, replay can be controlled through a foot pedal attached to the USB port on the computer. Figure 1 illustrates the set-up for transcription with a transcript window, the QuickTime video/audio window, and the Walker Controller window. The Walker Controller is set here to replay stretches of 4 seconds three times and then move on with a backspace of 1 second. It is also set here to play the media at 80% real speed.

During transcription, utterances are segmented based on criteria derived from syntax, intonation, pauses, and semantics, in accord with the analysis of Berndt, Wayland,
Rochon, Saffran, and Schwartz (2000). The CHAT transcription format includes conventions for marking linguistic behaviours such as word repetitions, revisions, fillers, gestures, sound fragments, and unintelligible output. Transcriptions are further elaborated by detailed coding for error type. This coding is done by certified, licensed speech-language pathologists with clinical and research experience in aphasia. For word-level errors we code errors in six categories: phonology, semantics, neologism, dysfluency, morphology, and formal lexical features (e.g., article errors). Within each
category word-level errors are coded further to indicate whether the error was a word or non-word, the target was known or unknown, a suffix was missing, and more. Errors that are not real words are transcribed using IPA. The error code also indicates if the error was repeated or retraced by the speaker within the utterance. The sentence-level codes capture empty speech, circumlocution, jargon, agrammatism and paragrammatism, and perseveration. A complete list of the error codes, definitions, and examples is available at the AphasiaBank website.

Checking

Every transcript goes through four levels of checking. The first level relies on the CHECK program that is built into the CLAN editor. To run this checker the transcriber types escape-L and looks to see if any errors are reported. We run this initial check several times during the production of a transcript. After initial completion transcripts are then reviewed by at least two transcribers for accuracy before being uploaded to the website. For the transcripts from PWAs one of those reviewers is always a certified, licensed speech-language pathologist. The third level of checking relies on part-of-speech tagging through the MOR program, which is described below. If a transcript can be run through MOR without errors we know that the transcript has no unrecognisable words. In effect, MOR serves as a filter for detecting misspellings and other incorrect lexical forms. The fourth level of checking relies on the Chatter program to test for complete adherence to the CHAT XML Schema.

All of the data in AphasiaBank have been run through all four levels of checking. Much of this work has been done at Carnegie Mellon, but we have also been able to train 14 transcribers at various sites to make reliable use of the CHAT coding system for AphasiaBank data. For researchers with training in language analysis and good familiarity with computers it takes about 3 days to learn to transcribe in CHAT for AphasiaBank data.

Morphological tagging

After completing transcription we conduct part-of-speech tagging using the MOR program. Analysis through MOR has two important functions. First, as we noted above, MOR acts as a filter against misspellings and other incorrect lexical forms. Because each word must have some recognisable morphological structure, MOR will catch and list many typos and other errors in transcription, which can then be corrected. Often transcribers wish to note that a word has a deviant phonological shape. At the same time they need to relate that shape to some standard word form. To do this CHAT uses a replacement form structure. For example, a deviant production of the word “pretzel” as “pezzle” can be transcribed orthographically or phonetically: pezzle [: pretzel] or pEz@l@u [: pretzel]. The MOR program will use the form in the replacement brackets, ignoring the preceding form. Another way of bridging the gap between actual productions and the forms seen by MOR is to enclose omitted material in parentheses, as in (be)cause for the word “because”. Additionally, transcribers can match speaker’s productions more closely to well-formed targets by using marks for repetition and retracing such as [/] and [//]. Faithful use of these and other transcription devices can greatly improve the quality of a transcription and facilitate the accurate running of MOR and POST.
The second function of MOR involves the use of the part-of-speech tags inserted by MOR for other programs. These tags provide a gateway to a wide variety of further automatic analyses of morphology, lexicon, and syntax. For the data currently in AphasiaBank, MOR analysis has been completed. However, it is important for users to understand how this analysis was constructed and how to interpret the tags.

We have developed MOR taggers for English, Spanish, German, French, Italian, Japanese, Cantonese, and Mandarin. For data currently in AphasiaBank we rely on the English MOR program. The results of the MOR tagger are then disambiguated using the POST statistical disambiguator (Parisse & Le Normand, 2000). POST uses the context before and after the word to assign part-of-speech to ambiguous cases. After running MOR, POST, and CHECK (to ensure that the output is complete and technically accurate), the transcript then appears with a new tier, %mor, under each speaker tier, giving the lexical and morphological tags for each word on the main speaker tier. These morphological codes can then be used to automatically compute a variety of indices and other linguistic analyses.

To judge the accuracy of tagging with MOR and POST we reviewed six control transcripts (over 20,000 total words) and did manual morphological tagging on a %trn line. Results demonstrated 98.87% agreement between the part-of-speech tagging done automatically by CLAN and that done manually. However, for new samples we expect accuracy to decline to about 98%, because statistical taggers always do better on training data than on new data. For transcripts from PWAs the level of tagging accuracy depends on the type of aphasia and other characteristics of the individual speaker. For participants with anomia or mild agrammatism, morphological tagging accuracy levels are close to those for controls. For participants with more severe agrammatism it is also possible to achieve high levels of accuracy for part-of-speech tagging, because the syntactic constructions being used are often quite simple, as in child language data where tagging of productions from children between 2 and 3 years of age has an accuracy level of about 96%. For participants with jargon aphasia it is often difficult to map productions onto standard word forms. For this group the tagging of word forms for the traditional part-of-speech categories is, by definition, less reliable. However, using the CHAT error coding system, the neologisms and word approximations of speakers with jargon aphasia can be classified systematically into various non-word categories and the remaining conventional words can be tagged with conventional tags.

The following example shows a few lines from a language sample from a participant describing his stroke, showing speaker lines (INV for investigator and PAR for participant) and their corresponding %mor lines. Some CHAT symbols that appear on the main speaker tier include: [ ] for repetition, [ // ] for revision, &= before gestures or simple events (e.g., &=laughs, &=sighs, &=sneezes), and & before sound fragments and fillers. On the %mor line, the part of speech (e.g., aux for auxiliary, pro for pronoun, v for verb) comes before the vertical bar and the word used by the speaker from the main tier. Suffixes are attached to the word (e.g., -PROG for progressive, -PAST for regular past).

*INV: can you tell me what you remember about it ?
%mor: aux|can pro|you v|tell pro|me pro|wh|what pro|you v|remember prep|about pro|it ?

*PAR: I remember falling off the chair and [ ] and &w &w &wonder &won wondering what happened to me.
%mor: pro|v|remember n|gerund|fall-GERUND prep|off det|the n|chair conj|coo|and n|gerund|wonder-GERUND pro|wh|what v|happen-PAST prep|to pro|me.
∗PAR: and I couldn’t get up & = laughs.
%mor: conj:coo|and pro|I aux|could~neg|not v|get adv:loc|up.
∗PAR: and I [//] it was morning.
%mor: conj:coo|and pro|it v:cop|be&PAST&13S n|morning.
∗PAR: and & uh & um it wasn’t until the afternoon that I called Alice.
%mor: conj:coo|and pro|it v:cop|be&PAST&13S~neg|not prep|until det|the n|afternoon rel|that pro|I v|call-PAST n:prop|Alice.
∗PAR: but I couldn’t say anything.
%mor: conj: coo|but pro|I aux|could~neg|not v|say pro:indef|anything.

Transcripts that have been tagged using MOR and POST can then be further analysed for syntactic structure using the GRASP Program (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010). This program takes the words on the %mor line and relates them in terms of a set of 34 binary syntactic dependency relations such as SUBJect, ADJunct, MODifier, and so on. The extraction of these relations allows researchers to study the use of syntactic patterns in aphasia and also supports the automatic computation of morphosyntactic profiles such as DSS (Lee, 1966) and IPSyn (Sagae, Lavie, & MacWhinney, 2005; Scarborough, 1990).

ILLUSTRATIVE ANALYSES

In this section we present a series of analyses of segments of the Cinderella story, designed to illustrate how we can analyse AphasiaBank transcripts. Our basic goal here is to illustrate the operation of the programs. At the same time, these analyses provide evidence regarding substantive issues in the study of aphasia. CLAN provides 15 analysis programs, each with a wide variety of functions and options. String-search programs can compute frequency counts, key-word and line profiles, mean length of utterance, mean length of turn, type-token ratios, maximum word length counts, maximum utterance length histograms, vocabulary diversity, temporal durations, and so on. The CHAT transcript files include header lines with several fields for demographic information, thereby making it possible to analyse subsets of the entire database based on, for example, sex or aphasia type.

MacWhinney et al. (2010) conducted a number of analyses of AphasiaBank data that focused on the segment of the protocol in which the participant retells the Cinderella story. Those analyses were done using the smaller database \( (n = 24 \) PWAs, \( n = 25 \) controls) that was available earlier in the project. Here we extend those analyses to the larger database currently available. The samples used for these analyses include all PWAs who meet these criteria: (1) aphasia caused by stroke; (2) relatively complete demographic and testing data; (3) native speakers of English; and (4) produced at least one utterance in response to the task. At the time of preparation of this manuscript, 90 PWAs met these criteria. We also selected 90 controls who met these criteria: (1) complete demographic data; (2) native speakers of American English; and (3) minimum age of 36.0 years. The characteristics of the PWAs and the controls appear in Table 1. Statistical tests revealed no significant difference between the two groups on the basis of age or education, but a chi-square test resulted in a significant difference, \( \chi^2(1) = 5.714, p < .05 \), between the groups on sex, with a higher ratio of males to females in the PWAs.

Before conducting the lexical frequency analyses, the Cinderella story segments of the transcripts were extracted using the GEM command. This command requires that segments of the transcript be demarcated with @G lines indicating the beginning of an
TABLE 1
Demographic and test data for AphasiaBank participants

<table>
<thead>
<tr>
<th></th>
<th>PWA, n = 90</th>
<th>Control, n = 90</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>64.5 years ($SD = 12.3$)</td>
<td>64.7 years ($SD = 13.8$)</td>
</tr>
<tr>
<td>range</td>
<td>range = 36–90.7</td>
<td>range = 36–87.8</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td>34 females</td>
<td>50 females</td>
</tr>
<tr>
<td></td>
<td>56 males</td>
<td>40 males</td>
</tr>
<tr>
<td>Handedness</td>
<td>77 right</td>
<td>80 right</td>
</tr>
<tr>
<td></td>
<td>7 left</td>
<td>8 left</td>
</tr>
<tr>
<td></td>
<td>4 ambidextrous</td>
<td>2 ambidextrous</td>
</tr>
<tr>
<td></td>
<td>2 unknown</td>
<td></td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td>15.7 years ($SD = 2.9$)</td>
<td>15.1 years ($SD = 2.4$)</td>
</tr>
<tr>
<td>range</td>
<td>range = 12–25</td>
<td>range = 10–22</td>
</tr>
<tr>
<td><strong>Time post-onset</strong></td>
<td>6.6 years ($SD = 6.0$)</td>
<td></td>
</tr>
<tr>
<td>range</td>
<td>range = 0.5–39.2</td>
<td></td>
</tr>
<tr>
<td><strong>Type of aphasia</strong></td>
<td>32 anomic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 Broca</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16 conduction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9 Wernicke</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 not aphasic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 transcortical motor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 global</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 unavailable</td>
<td></td>
</tr>
<tr>
<td><strong>WAB AQ (aphasia quotient)</strong></td>
<td>71.8 ($SD = 18.9$)</td>
<td></td>
</tr>
<tr>
<td>range</td>
<td>range = 17–97.6</td>
<td></td>
</tr>
</tbody>
</table>

activity. In this case, @G: Cinderella Story is used to mark the beginning of the participant’s retell. In addition, it is necessary to exclude extraneous comments made during the task, such as “wait a second”, “can’t say it right”, and “I know”. Those utterances are marked with a sentence level code [+ exc] during transcription. We typed the following command into the CLAN command window to extract the Cinderella retell portion of the transcript.

gem +sCinderella +sStory -s"[+ exc]" +g +n +d1 +t*PAR +t%mor +f +re *.cha

This command has 11 segments:

gem calls up the GEM command
+sCinderella +sStory searches for the words Cinderella and Story
-s"[+ exc]" excludes utterances coded for exclusion
+g selects the @G or @BG segment that has all and only the words specified by the +s option
+n ends the segment at the next @G in the transcript
+d1 creates output in legal CHAT format
+t*PAR includes the participant speaker tier
+t%mor includes the %mor tier
+f sends output to a new file (with .gem.cex extension)
+re runs program subdirectories within a folder
+.cha runs the command on all the .cha files
Lexical frequency analysis

Once we had extracted the relevant segments of the transcript using GEM, we could then proceed with further analyses. Our first set of analyses uses the FREQ program to compute these four forms of lexical frequency analysis: (1) overall sex and group differences, (2) top 10 item differences, (3) noun usage, and (4) verb usage.

Sex and group difference analysis. The first FREQ analysis examined the relative distribution of word forms across participant group and sex. To get the total number of words (tokens) and the number of different words (type), excluding repetitions, revisions, and unintelligible words, the following FREQ command was used.

freq +t*PAR -r6 +re +d3 -sxx *.gem.cex

The five new components of this command are:

freq calls up the FREQ command
-r6 excludes repetitions and revisions
+d3 outputs type/token information for analysis in Excel
-sxx excludes unintelligible words
*.gem.cex runs the command on all files with the .gem.cex extension

This command can be modified to look specifically at only females (or males) by adding one more element: +t@id=""|female|"" (or +t@id=""|male|"").

Results of the type and token analysis for both PWAs and controls appear in Table 2. One can see that the controls produced more than twice as many total words and different words as the PWAs, \( t(178), p < .001 \), one-tailed. While one might imagine that females might produce more output than males in their Cinderella story retells, we found no evidence to suggest that was the case. Differences between males and females were not significant in the controls for total number of words, \( t(88), p = .69 \), or total number of different words, \( t(88), p = .60 \). Differences between males and females were more pronounced in the PWAs, although they did not reach statistical significance for total number of words, \( t(88), p = .08 \), or total number of different words, \( t(88), p = .06 \).
Top 10 analysis. The second FREQ analysis examined the overlap between groups in terms of the top 10 words used in describing the Cinderella story. MacWhinney et al. (2010) used FREQ to conduct a detailed comparison of the top 10 words in the PWA samples with the top 10 in the control samples. Based on the sample at that time of 24 PWAs and 25 controls, they showed a nearly complete overlap between these lists for the top 10 words, although PWAs had a markedly reduced lexical diversity. Here we extend this analysis to our larger sample of 90 PWAs and 90 controls. The first analysis computes the frequencies of word form occurrences on the %mor line, excluding neologisms and unintelligible words. In cases where PWAs made an error, but the intended word was known (e.g., phonological errors in the pronunciation of “Cinderella”, semantic errors such as “foot” for “shoe”), the intended word was used for this analysis. The command we used to construct the overall frequency profile was:

freq +t%mor +t*PAR -t* +s+r,-o-,% +u -s@"|-neo,|-unk" +o +f +re *,gem.cex

From the output produced by this command we extracted the occurrences of the 25 most-frequent words. The first two columns of Table 3 provide these 25 most-frequent words across all parts of speech for the PWA and control groups’ stories.

The previous article reported the top 15 words for the PWAs only. The current larger sample yielded nearly the same list in roughly the same order. Three words—
then, have, and so—appeared in the top 15 for this larger group, displacing three—not, he, and I—from the previous list. In the current analysis the top 25 words are similar for the PWAs and controls with the exception of 5 words: then, but, I, say, and oh appear in the aphasia list; in, prince, all, slipper, and with appear in the list for the control participants. Overall, this analysis replicates the results of MacWhinney et al. (2010).

**Noun usage.** We also examined the frequencies of nouns across the PWA and control groups. To do this, we used the following command to tabulate the frequencies of all nouns, including proper nouns, compound nouns, and nouns with prefixes, collapsing across stems:

```
freq +t%mor +t*PAR –t∗+o+s@r∗,|n∗,|n,o-% +u +re *.gem.cex
```

The results of this analysis are given in Columns 3 and 4 of Table 3. PWAs produced about half as many (383 vs 661) different nouns as did the controls with about a third the number (3062 vs 8289) of noun tokens (total noun stems). The two lists of most frequently occurring nouns have 20/25 words in common (slightly higher than the 6/10 reported in the previous study). Again, the stories from the PWAs included the words man, thing, person, o'clock, and sudden, which are not as tightly and specifically linked to the Cinderella story as are the words midnight, carriage, foot, father, and castle that appear in the controls’ top 25. Additionally, in the PWA samples, girl, another less-specific word, is the 3rd most frequently occurring noun, as opposed to being 13th in the non-aphasic samples. And the word glass, which is used to describe a detailed aspect of the slipper in the Cinderella story, was 23rd on the list of word frequency for the PWAs but 7th on the controls’ list. Interestingly, although the PWAs used the word o’clock only 41 times, they used twelve, in the combination twelve o’clock, 35 out of these 41 times. They used the word twelve 52 times in all (17 times without the word o’clock immediately following) and they used the word midnight 19 times. The controls said midnight 126 times, twelve o’clock together as a unit only 18 times, twelve 53 times, and o’clock 19 times.

**Verb usage.** As was the case for nouns, PWAs produced just over half as many (317 vs 590) different verbs as did the controls with just over a third as many total verbs stems in the sample (3657 vs 9371). We used the following command to find all verbs, auxiliaries, and participles, again collapsed across stems:

```
freq +t%mor +t*PAR –t∗+o+s@r∗,|n∗,|n,o-% +u +re *.gem.cex
```

Because it can be difficult for users to remember how to construct complex commands of this type, we have provided a user interface for the construction of the +t and +s switches, as illustrated in Figure 2. The top 25 most frequently occurring verbs in both groups appear in columns 5 and 6 of Table 3. In this case the part of speech label is included to indicate the exact form of the verb (e.g., verb, copula, verb participle, auxiliary). The most frequently occurring verbs have 21/25 words in common across the two groups (higher than the 7/10 reported in the previous study). The stories from the PWAs included more frequent use of the verbs think, dance (in verb and participle forms), and look, whereas the stories from the controls included more frequent use of the verbs turn, live, try, and marry.
Lexical diversity analysis

One possible consequence of aphasia is a reduction in lexical diversity (Wright, Silverman, & Newhoff, 2003). For example, we may be interested in understanding whether individuals with non-fluent aphasia have a greater reduction in lexical diversity than do individuals with fluent aphasia. Traditionally this feature of aphasic speech has been measured using the type-token ratio (TTR) measure (Holmes & Singh, 1996). However, a major problem with TTR is the fact that it is overly sensitive to sample size, because frequent words only demonstrate their impact in larger samples. Particularly for short samples from individuals with non-fluent aphasia, TTR could provide an inflated estimation of lexical diversity. To address this problem, Malvern, Richards, Chipere, and Purán (2004) developed the VOCD (Vocabulary Diversity) statistic. VOCD corrects the problem with TTR by selecting up to 20 alternative sample sizes for computation of the type ratio. This then allows the program to plot and compare the lexical diversity function independently of some particular sample size. Because VOCD is superior to TTR in this regard, Table 2 reports VOCD scores, rather than TTR scores. One can compute VOCD either from the main speaker line or the %mor line in the CHAT transcript. The advantage of calculating VOCD from the %mor line is that one can do an analysis that focuses on lemmas, rather than word forms. That is, we can treat variant inflected forms of the same base (e.g., play, playing, played or unhappy, happy, happily) as the same lexical item, thereby obtaining a more accurate measure of lexical diversity. The following command was used to do this analysis, excluding neologisms and unintelligible words as well.

```
vocd +t%mor +t*PAR -t* +s"*" +s"*" +s"*" +s"*" -&%" -s@" -neo, |-unk" +d3 +re*.gem.cex
```

Results of the VOCD analysis also appear in Table 2. VOCD could not be computed in 11 PWA samples and one control sample, because there were not enough tokens for random sampling without replacement. VOCD differences between males and females were not significant for either group but the VOCD difference between groups was significant, $t(178)$. $p < .001$, with PWA samples including about 60% of the lexical diversity seen in the control samples.
Morphosyntactic analysis

Researchers have proposed several systems for analysing control of morphosyntactic markings and patterns in normal and disordered speech. Among the most well-known systems are LARSP (Fletcher & Garman, 1988), DSS (Lee, 1966), IPSyn (Scarborough, 1990), and QPA (Rochon, Saffran, Berndt, & Schwartz, 2000). Because each of these systems requires careful hand analysis, tagging, and coding, they have only been used sparingly in research studies and clinical practice. CLAN’s MORtable program provides automatic computation of many of the indices that figure prominently in these earlier hand-coded systems.

The MORtable command works automatically on the %mor line to construct a table of the parts-of-speech that can be opened directly in Excel. The command has this form:

```
mortable +t*PAR +u +re *gem.cex
```

The results appear in Table 4. Looking at the right-hand column with proportions, one can see that the PWA samples were disproportionately low in these parts of speech: reflexive pronouns, possessive pronouns, prepositions, modals, and infinitives. For the bound morphemes that mark regular inflectional morphology (MacWhinney, 1978; Pinker, 1979), the PWA samples were disproportionately low in the use of the superlative, possessive, regular third person singular, and regular past tense.

The output of MORtable can serve as the input to automatic computation of indices such as LARSP, DSS, IPSyn, and QPA. Currently, CLAN provides this type of automatic computation for DSS and IPSyn.

Apart from the across-the-board analyses provided by MORtable, researchers can also conduct targeted analyses of morphosyntactic structures using the COMBO program. COMBO is designed to search for syntactic and collocational patterns with variables across either the main line or the %mor line. In the Cinderella story material extracted by GEM, we searched the groups for the use of these collocations: *once upon a time*, *happily ever after*, *glass slipper(s)*, and *fairy godmother*. The following COMBO commands were used, using the +r6 option to exclude revisions and repetitions within utterances.

```
combo +t*PAR +re +u +r6 +sonce`upon`a`time *.gem.cex
combo +t*PAR +re +u +r6 +shappily`ever`after *.gem.cex
combo +t*PAR +re +u +r6 +sfairy`godmother *.gem.cex
combo +t*PAR +re +u +r6 +sglass`slipper* *.gem.cex
```

In the PWA samples, *once upon a time* did not occur at all; in the control samples it occurred 13 times. *Happily ever after* occurred only 10 times in the samples from PWAs and 61 times in the control samples. *Glass slipper(s)* occurred together 20 times in the samples from the PWAs and 199 times in the control samples. *Fairy godmother* occurred together 26 times in the PWA samples and 153 times in the control samples. The relative infrequency of these collocations in the PWA samples seems to indicate a diminished use of finer levels of narrative expression. Interestingly, in the PWA samples (excluding revisions and repetitions) the word *fairy* was used 41 times and the word *godmother* appeared 38 times, illustrating that sometimes only one of the words was used and sometimes the words were both produced in the sentence but not as an uninterrupted unit. In the control samples, *fairy* occurred 172 times and *godmother* 169 times, much closer to the 161 times these two words occurred together. Likewise,
### TABLE 4
Parts of speech and bound morpheme frequency counts

<table>
<thead>
<tr>
<th>Parts of speech</th>
<th>PWA (90)</th>
<th>Control (90)</th>
<th>PWA/Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>wh-words</td>
<td>105</td>
<td>236</td>
<td>.44</td>
</tr>
<tr>
<td>adjectives</td>
<td>667</td>
<td>1621</td>
<td>.41</td>
</tr>
<tr>
<td>adverbs</td>
<td>1244</td>
<td>2770</td>
<td>.45</td>
</tr>
<tr>
<td>auxiliaries</td>
<td>442</td>
<td>1022</td>
<td>.43</td>
</tr>
<tr>
<td>complements</td>
<td>0</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>conjunctions</td>
<td>2155</td>
<td>4243</td>
<td>.50</td>
</tr>
<tr>
<td>determiners</td>
<td>1759</td>
<td>4479</td>
<td>.39</td>
</tr>
<tr>
<td>infinitives</td>
<td>311</td>
<td>1022</td>
<td>.30</td>
</tr>
<tr>
<td>modals</td>
<td>142</td>
<td>472</td>
<td>.30</td>
</tr>
<tr>
<td>nouns</td>
<td>3062</td>
<td>8289</td>
<td>.37</td>
</tr>
<tr>
<td>negatives</td>
<td>205</td>
<td>410</td>
<td>.50</td>
</tr>
<tr>
<td>prepositions</td>
<td>955</td>
<td>3493</td>
<td>.27</td>
</tr>
<tr>
<td>pronouns</td>
<td>2370</td>
<td>5062</td>
<td>.47</td>
</tr>
<tr>
<td>possessive pronouns</td>
<td>131</td>
<td>628</td>
<td>.21</td>
</tr>
<tr>
<td>reflexive pronouns</td>
<td>6</td>
<td>37</td>
<td>.16</td>
</tr>
<tr>
<td>quantifier, determiner:number</td>
<td>454</td>
<td>782</td>
<td>.58</td>
</tr>
<tr>
<td>verbs, copulas, participles</td>
<td>3073</td>
<td>7877</td>
<td>.39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bound morphemes</th>
<th>PWA (90)</th>
<th>Control (90)</th>
<th>PWA/Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd person singular – regular</td>
<td>247</td>
<td>1191</td>
<td>.20</td>
</tr>
<tr>
<td>3rd person singular – irregular</td>
<td>476</td>
<td>807</td>
<td>.59</td>
</tr>
<tr>
<td>past tense – regular</td>
<td>203</td>
<td>743</td>
<td>.27</td>
</tr>
<tr>
<td>past tense – irregular</td>
<td>1023</td>
<td>2059</td>
<td>.49</td>
</tr>
<tr>
<td>comparative</td>
<td>11</td>
<td>22</td>
<td>.50</td>
</tr>
<tr>
<td>superlative</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>plural – regular</td>
<td>447</td>
<td>1147</td>
<td>.39</td>
</tr>
<tr>
<td>plural – irregular</td>
<td>81</td>
<td>215</td>
<td>.37</td>
</tr>
<tr>
<td>possessive</td>
<td>17</td>
<td>119</td>
<td>.14</td>
</tr>
<tr>
<td>perfect participle – regular</td>
<td>95</td>
<td>261</td>
<td>.36</td>
</tr>
<tr>
<td>perfect participle – irregular</td>
<td>47</td>
<td>187</td>
<td>.25</td>
</tr>
<tr>
<td>progressive</td>
<td>318</td>
<td>763</td>
<td>.41</td>
</tr>
</tbody>
</table>

the word *glass* occurred 31 times and *slipper(s)* 100 times in the samples from the PWAs; and 217 times and 365 times, respectively in the control samples.

**Error analysis**

To capture word-level and sentence-level errors we apply an extensive coding system that has been shaped specifically for AphasiaBank transcripts. Using FREQ, one can search for variant forms of production of a word such as *Cinderella* with the command below:

\[
\text{freq } +t^\ast\text{PAR } +s^\ast\text{Cinderella}^\ast\ast^\ast^\ast^\ast +d6 +re +u^\ast\ast\ast\ast\ast\ast\ast\ast \text{.gem.cex}
\]

Results reveal that the word *Cinderella* (and *Cinderella’s*) was produced by the PWAs 349 times (including repetitions and retracings), and 23% (79) of those were not correct. In only six cases was an attempt made to revise the incorrect production within the same sentence. In no cases was the same error repeated within the same sentence. Some of the errors met the strict criteria used in this project for phonological
errors, but most did not. For an error on a multi-syllabic word to be coded as a phonological error, the error must have complete syllable matches on all but one syllable, and the syllable with the error must match on two out of the three elements of the syllable (onset, vowel nucleus, coda). So, for example, Cindewella and Cinberella would be considered phonological non-word errors (coded [∗ p.n]), but Cillewilla, Cellerella, and Swinwella would be considered neologisms with a known target (coded [∗ n.k]). (All non-word errors in the transcripts are transcribed in IPA.) A list of all Cinderella error productions by error type is given in Table 5.

A variety of interesting error analyses could be conducted, for example, looking at related versus unrelated semantic paraphasias, the number of phonological errors in neologisms for known targets, the number of errors that are repeated, the ways in which errors are revised, errors on particular parts of speech such as pronouns, proportions of errors in free speech versus picture descriptions, and errors by type or severity of aphasia. We will report on a few of these for purposes of illustration. Using the command below, a list of related and unrelated semantic errors was generated.

freq +s["\s:t"] +s["\s:s"] +u +re +t*PAR +d6 *.gem.cex

Results revealed 230 semantically related errors and 10 semantically unrelated errors in the Cinderella stories by the PWAs. In 80 cases attempts were made to revise the error within the same utterance, and in 8 cases the error was repeated within the same utterance. Examples of the unrelated semantic errors are: trucks for dress, weather for Cinderella, words for wand, building for carriage, and bliss for slipper. The most commonly occurring related semantic errors are he for she, she for he, him for her, his for her, foot for shoe, mother for stepmother, and princess for prince.

### Table 5

Error productions of Cinderella (with frequency if greater than 1)

<table>
<thead>
<tr>
<th>Phonological (non-word), N = 31</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sindwella (6)</td>
<td>twindwella</td>
</tr>
<tr>
<td>sindwella (5)</td>
<td>sindwella</td>
</tr>
<tr>
<td>kinde-swa (3)</td>
<td>sindwella</td>
</tr>
<tr>
<td>sindwella</td>
<td>sindwella</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neologism (known target), N = 41</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>s沿海一l</td>
<td>s沿海一l</td>
</tr>
<tr>
<td>s沿海一l</td>
<td>s沿海一l</td>
</tr>
<tr>
<td>s沿海一l</td>
<td>s沿海一l</td>
</tr>
<tr>
<td>s沿海一l</td>
<td>s沿海一l</td>
</tr>
<tr>
<td>sindwella (3)</td>
<td>sindwella</td>
</tr>
<tr>
<td>sindwella (3)</td>
<td>sindwella</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantic related, N = 5</th>
<th>Cindy</th>
<th>Cin (2)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Semantic unrelated, N = 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>weather (2)</td>
<td></td>
</tr>
</tbody>
</table>
The preponderance of pronoun for pronoun substitutions can be investigated further using the command below to generate a list of pronouns and errors.

freq +%mor +t*PAR -t* +d6 +o +s"[@r-.|]*pro,o-%" +s"[@r-.|]*pro:,.o-%" +re +u *.gem.cex

Errors occurred on 86 of 2579 pronoun productions (excluding repetitions and revisions) in the aphasia stories. In 72 of those cases (84%) the error was another pronoun; in 1 case the error was in for it, which could be considered phonological or semantic, but still not another pronoun; 6 errors were purely phonological in nature (e.g., see for she, bay for they); 4 pronouns were produced with unknown targets (referents); 2 were produced dysfluently (with syllable insertions); and 1 error was phonological in nature but was coded as a neologism because it had multiple element changes (herfers for herself).

Another form of error analysis tracks the ways in which PWAs attempt to produce the collocations most relevant to the story. Earlier, we saw that collocations like glass slipper, happily ever after, and fairy godmother were relatively rare in the PWA samples. When we look at the productions of these forms, we see that they are often produced as errors. For example, glass slipper(s) were called glass skipper, glass sippers, glass ball, and glass crystal. Fairy godmother appeared as fairy god, fairy grandmother, fairy goymother, fairy mother, grairy godmother, sairy godmother, and firey godmother. Examples of some paraphasic errors for happily ever after appear below.

*PAR: they live hevry [: happily] ever after.
*PAR: &uh and they’re &maf hæf I pl I @u [: happily] ever after.

While very few PWAs managed to say happily ever after, many of them attempted to communicate the concept at the end of their stories in a variety of alternative ways, some of which are shown below.

*PAR: he’s very happy with her.
*PAR: and a &h happy life &=laughs.
*PAR: and so &um the prince and &uh Cinderella &i <is no> [/] was &uh very &h happy all [/] all the way through.
*PAR: yes but &uh got@u and Cinderella hv@u [: live] happy together.
*PAR: and &um &um the man and the [/] &uh stnd@r@r-dv@u [: Cinderella] olwet@u [: always] &b happy ever æf@r@u [: after].
*PAR: and then he [/] she is &um &uh in the end had a very in happy &w &uh place.
*PAR: well she wears [: lives] &ev happy of ever.
*PAR: &hap happy [/] happy something yeah.
*PAR: and &uh &a &=fingers:writes æpt@u [: happy] [/] æpt@u [: happy] &=hand:no no [/] &uh &ch and &eh marriage.
*PAR: &=sighs &hm (.) &uh &=sighs I think &=ges:unsure Cinderella’s hæv@u [: happy] ever after.
*PAR: and so she’s happy ever after.
*PAR: <and they> [/] &ha [x 3] and they (.) happy ever after.
*PAR: and so the (.) married and [/] &=shrugs and &uh married &h after.

Gesture analysis
There have been a number of analyses of the use of gesture by people with aphasia in spontaneous communicative situations (Glosser, Wiener, & Kaplan, 1986; Goodwin, 2000, 2003a, 2003b). These analyses suggest that gesture can compensate to some
degree for verbal deficits (Fex & Månsson, 1998). However, the nature of this compensa-
tion varies markedly across aphasia types (Cicone, Wapner, Foldi, Zurif, & Gardner, 
1979). Because AphasiaBank dialogues are collected in the same fashion across tasks 
we can make consistent assessments of the usage of gesture within and across these 
types.

The detailed analysis of gestural patterns can be very time consuming. Many analy-
ses of gesture have focused on the nature of the synchrony between speech and gesture 
(A llen et al., 2007; McNeill, 1985). For analyses of this type, investigators often rely 
on linkage of speech to gesture through the Elan (http://www.lat-mpi.eu/tools/elan/) 
video annotator. The CLAN program called CHAT2ELAN converts AphasiaBank 
transcripts to Elan. Researchers can then analyse gesture–speech patterns inside 
Elan. Afterwards they can export the data back to CHAT, using the ELAN2CHAT 
program.

CLAN also provides its own methods for analysing gestures. These methods focus 
less on the synchrony between speech and gesture (although that can also be anno-
notated) and more on the profiling of gestural sequences, as analysed in Kendon (1982).

To briefly illustrate this approach, consider the sequence of gestures produced by par-
ticipant adler11a. This 80-year-old male participant has severe Broca’s aphasia (WAB 
AQ = 17) and is 12 years post-onset. Within the context of the AphasiaBank proto-
col this participant produced only the words oh, no, well, and hello. Yet, through his 
gestures, he was able to communicate about the following 19 events within the context 
of the retelling of the Cinderella story: shaking hands (indicating the start of the nar-
"rative), turning a page in the book (moving on to the next event), stepsisters playing 
(accompanied by singing to illustrate play), Cinderella sweeping (brushing on table), 
turns page, receiving invitation to ball (hand receives paper), Cinderella requesting 
to go (wanting expression and fingers indicating walking), stepmother refusing (head 
shake and word “no”), clothing selection (hands touch objects, followed by “well”), 
looking beautiful (smiles, looks up, and gasps at beauty), ripping of the dress (cross-
body arm motions), throwing the dress away (tossing object across body and then 
forcefully to side), crying (head on arm, hand covers face, sobs), turns the page, being 
at the ball (dancing gestures with hands and body, sings tune from Disney movie), 
turns page, falling in love (gasps, hand to heart, head down, hugs), losing a shoe 
(object drops and hand reaches under table, gestures object slipping away), kissing 
goodbye, Prince knocks on door, waving hello, and kissing again. This entire sequence 
of 19 events was produced in a span of 49 seconds.

Figure 3 illustrates how these activities are coded in the CLAN transcript with a 
focus on the activities at line 721 describing the grabbing and ripping of Cinderella’s 
dress. In the right hand side of this screenshot we see the basic CHAT transcript with 
codes like &=imit:ripping describing the ripping action (and sound effect) produced 
by the participant. In the bottom left we see the QuickTime video window, which 
can be controlled through links from the transcript or by the scroll buttons in that 
window. In the top left we see a text file called 5dress.cut, which analyses the sequence 
of gestures beginning at line 718. This breakout window focuses on the analysis of the 
three segments of fifth gesture sequence, which are labelled as 5A-5B-5C-5D-5E-5F-
5G-5H. Only the first two segments of the sequence are visible in the part of 5dress.cut 
shown here, but the others follow below those. Each gesture is then further analysed 
to indicate the dynamics of the action, the major body part involved, the classification 
of the gesture, and its functional meaning. The classification field uses keywords that 
can then be searched with FREQ.
Some of these gestures are produced as single gesture sequences with separate retraction and some are produced as parts of larger sequences. For each of these gesture sequences we then create a small separate small file link to the main transcript file. Here is an illustration of the breakout for the sequence in which the stepmother rips off and discards Cinderella’s dress. The main transcript has this information:

```
*PAR: & = head:turns & = breath:in & = imit:ripping & = takes & = ges:away
& = imit:crying & = hand:flip okay.
@G: dress ripping sequence •%txt:”5dress” •
```

When the transcriber clicks on the bullet at the end of the @G line, a secondary file called 5dress.cut opens up. This file allows the transcriber to enter freeform coding of the gesture sequence. Here is a sample of the first few lines of this file:

```
@Media: adler11a, movie
Sequence: 5A-5B-5C-5D-5E-5F-5G-5H-5J-5K
Segment: 5A
```
This file then continues with an analysis of the remaining eight segments of the dress-ripping sequence. When this file is opening up directly from the main file it displays the clips associated with the first frame of each segment given in the %pic lines. In general, this form of analysis allows us to understand the details of the ways in which this participant uses gesture to convey a rich understanding of the Cinderella story.

Phonological and temporal analyses

AphasiaBank data can be analysed for phonological and phonetic structures using the Phon program (http://childes.psy.cmu.edu/phon) that was developed to be compatible with CHAT and CLAN. Phon use IPA Unicode on the %pho line in CHAT files to encode the precise phonological form of utterances. Using this information it can also perform automatic syllabification and alignment with target phonological structures. The audio corresponding to individual utterances in CHAT can be sent to Praat (www.fon.hum.uva.nl/praat/) for further detailed phonetic analysis and the results can then be recaptured in Phon/CHAT files.

CLAN also provides the TIMEDUR program that can compute overall session length, the durations of individual utterances and the lengths of the pauses between utterances. Programs such as DIST, COOCCUR, KEYMAP, and CHAINS can be used to analyse various aspects of sequences in discourse structure.

FUTURE DIRECTIONS AND CONCLUSION

The goal of this paper has been to illustrate the ways in which AphasiaBank data can be used to address substantive issues in the study of aphasia. We have shown how many indices and analyses that were previously computed by hand can now be computed automatically and accurately using the CLAN programs. Moreover, unlike previous work, the data, procedures, and results for these analyses are now being made fully public and can therefore be replicated and even challenged by anyone in the scientific community. The illustrative studies presented here constitute only a small sampling of the studies that can be conducted with these data. Additional ideas for future studies using the AphasiaBank database are continually being generated and posted at a link on the home page of the AphasiaBank website.

As AphasiaBank moves into the future it will need to confront several interesting challenges. As we move to collect data from a wider variety of languages we will need to translate and revise the protocol to match local cultural expectations, as well as patterns of bilingualism. We will need to develop specific methods for best assessing the language abilities of participants with global aphasia. We will need to implement data
collection procedures that will maximise our ability to include naturalistic conversational data in settings such as meal preparation, event planning, game playing, and other casual interactions. We will also need to apply AphasiaBank methods to study the impact of therapy treatments. Finally, as we move forward with this expansion of the database and the scope of AphasiaBank we will want to build new programs and new methods of analysis. Fortunately we can address these future challenges with confidence, knowing that we have already constructed an important new tool for the study of aphasia.

REFERENCES


