AphasiaBank: Data and Methods

Brian MacWhinney, Davida Fromm, Audrey Holland, and Margie Forbes

Carnegie Mellon University

14.1 Introduction

Recent years have seen a rapid growth in the use of computerized databases throughout the sciences. In the area of language studies, many of these databases involve the collection of large amounts of either spoken or written language. These collections, called corpora, are then accessed over the Internet and subjected to a variety of analyses for language structure, function, and content. The conditions of access to these corpora vary widely. However, some corpora, such as those in the CHILDES or TalkBank databases, are freely open to all researchers.

Typically, individual corpora have been collected and organized with a specific set of research questions in mind. For example, the CHILDES database includes corpora from children growing up in bilingual families in which care has been taken to collect the two languages when used either separately or together (Yip & Matthews 2007). Other corpora may focus on the use of language in classrooms (Goldman et al. 2007) or doctor-patient interaction (Frederiksen et al. 2004).

Corpora can be subjected to a wide variety of analyses. One can study changes in language over time during language learning (Brown 1973) or recovery (Feldman et al. 1994). Corpora may be sampled systematically across different social situations or levels (Labov 2001) or using different elicitation formats (Schober & Conrad 2006). The analyses can be conducted through methods as diverse as computer modeling (MacWhinney &

Leinbach 1991), microgenetic analysis (Siegler 2006), Conversation Analysis (MacWhinney & Wagner 2010), and computational linguistic methods for automatic grammatical analysis (MacWhinney 2008).

14.2 AphasiaBank

Some research areas have made more extensive use of corpora than others. One area that has benefitted particularly from the availability of open-access corpora has been the field of child language research. In this area, over 3,500 articles have been published using the CHILDES database (http://childes.psy.cmu.edu). CHILDES is an international cooperative venture, involving some 3,000 users 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 (MacWhinney 2010). The system provides users access to a set of programs (CLAN), a database (CHILDES), a transcription system (CHAT), documentation, and a mailing list (Info-CHILDES) for communicating on problems in language analysis. The form of these tools has been shaped by continual input from active members of the system.

The AphasiaBank Project seeks to extend the methods and procedures developed in the CHILDES system to the study of language in aphasia. To achieve this, the AphasiaBank project has developed a shared database of multimedia interactions for the study of communication in aphasia. This database now provides both a powerful platform for improving our understanding of aphasia and its treatment. In this chapter we will describe the goals of the project, the process of development, and the various analyses that are being conducted on the data. These analyses will illustrate a wide-ranging set of new tools for the analysis of language production in aphasia. By improving access to a shared database on aphasia, we can achieve a rapid improvement in the empirical grounding of work in this field. Together, the new database and the new analytic system will be able to support a major revolution in this field.

The organization of AphasiaBank began with a planning meeting of 20 senior aphasia researchers in 2005 who agreed on the need for a shared protocol, a shared database, and increased availability of computational tools for the aphasia research community. After funding was awarded in 2007, we developed the testing protocol and began collecting data from research and clinical aphasia centers around the country. To date, the database includes over 120 participants with aphasia from 10 different sites and just over 100 non-aphasic adults from 3 sites. The core database has been limited (with few exceptions) to individuals whose aphasia results from a stroke that can be verified through neuroimaging or a clear medical diagnosis. The current samples are all in English, with additional new samples being collected in Cantonese, Mandarin, German, and Swedish.

Section Summary: Electronic corpora that are openly available over the Internet are providing an increasingly powerful research tool for various areas in language studies. Recently, the AphasiaBank Project has extended these methods to the study of aphasia.

14.3 Goals

The overarching goal of work in AphasiaBank is the construction of methods for improving patient-oriented treatments in aphasia. To reach that goal, we must solidify the empirical database supporting our understanding of communication in aphasia. The consortium of aphasia researchers that has been involved in this project throughout its inception and development has continued to contribute to a shared conceptual and methodological framework that drives the collecting, recording, transcribing, and coding of language samples. The nine specific aims of AphasiaBank are:

1. **Protocol standardization.** We have developed a standardized data collection protocol that is being implemented at all consortium sites. Use of this standardized protocol guarantees maximal comparability across data sets.
2. **Database development**. We are compiling data from a large number of participants. Transcripts are done using the CHAT system (MacWhinney 2000) and linked to the digitized audio and video.
3. **Analysis customization**. Using the current CLAN programs (MacWhinney 2000) as a basis, we have constructed a set of 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 measures that were otherwise being coded by hand. We are also developing new measures based on automatically constructed annotations.
5. **Syndrome classification**. Using these new measures and the growing database, we are working with consortium members and statistical consultants to develop new approaches to syndrome-based patient classification and diagnosis.

6. **Support for qualitative analysis.** We are supporting qualitative analysis on three levels. First, the CLAN editor supports standard Conversation Analysis (CA) transcription. Second, we have formalized a set of coding systems specific to communications involving persons with aphasia. Third, we are promoting a system to provide web-based collaborative commentary on conversational interactions.
7. **Characterization of recovery processes**. We will develop microgenetic methods such as time sequential analysis and growth curves to trace changes across time in both individual participants and groups of participants.
8. **Evaluation of Treatment Effects**. We will develop methods that allow us to evaluate the effectiveness of specific aphasia rehabilitation treatments and are beginning to get repeated measures at yearly intervals from some participants.
9. **Johnny Appleseed**. We are disseminating these new tools through personal contact, annual workshops, journal publications, conference presentations, the AphasiaBank Google Group, and downloads available over the Internet.

Section Summary: The goals of AphasiaBank are protocol standardization, database development, analysis customization, measure development, syndrome classification, support for qualitative analysis, characterization of recovery processes, evaluation of treatment outcomes, and dissemination of methods.

14.4 Protocol Standardization

Based on extensive input from consortium members and pilot work, we have established a uniform AphasiaBank protocol. This protocol, along with demographic forms, demographic spreadsheets, tests, and stimuli are all available at the AphasiaBank website (http://www.talkbank.org/AphasiaBank). The protocol consists of four different discourse genres: personal narratives, picture descriptions, story telling, and procedural 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 ten seconds. A troubleshooting script is also available for participants who still cannot respond and need additional prompting with simplified questions. The discourse protocol is administered in one session and is recorded on video. The investigator makes every effort to be as silent as possible while giving maximal non-verbal encouragers to the participants. Participants are given as much time as they need for their responses.

The personal narratives are elicited by asking the participants with aphasia about their speech, their stroke, their recovery, and an important event in their lives. Non-aphasic 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.

For the 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. The first picture stimulus is a four-paneled picture of a child playing with a soccer ball and breaking a window, the second is a six-paneled picture of a child refusing an umbrella and getting caught in the rain, and the third is the Nicholas and Brookshire (1995) picture of a cat stuck in a tree. A fourth picture, a color photo of a flood rescue scene, was used for the first two years of the project and then discontinued because many participants were having trouble interpreting the picture.

For the story telling task, participants are shown a paperback picture book of *Cinderella*, with the words covered. 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.

Finally, the procedural discourse task involves asking the participants to describe how they would make a peanut butter and jelly sandwich. (Test sites outside the United States may substitute another simple food preparation.) A stimulus picture with photographs of peanut butter, bread, and jelly is available for use with participants who need extra help.

In addition to the discourse protocol, four tests are administered to participants with aphasia: 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 et al. 2001); 3) the Verb Naming Test from the Northwestern Assessment of Verbs and Sentences-Revised (Thompson, in preparation); and 4) the AphasiaBank 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 non-aphasic participants are tested with the Mini-Mental State Exam (Folstein et al. 2002) and the Geriatric Depression Scale (Brink et al. 1982) to rule out cognitive impairment and depression. All test results are entered into a master spreadsheet that is password protected on the AphasiaBank website and available to AphasiaBank members.

Finally, in addition to the discourse protocol and the testing, investigators collect extensive demographic information about all participants. Fifty-one fields of data in the demographic spreadsheet include variables such as gender, date of birth, race, handedness, education, occupation, language status (monolingual, childhood bilingual, etc.), aphasia etiology, aphasia duration, aphasia type, site of lesion, motor status, depression, dysarthria, apraxia of speech, history of neurological conditions, and history of communication disorders. Table 1 provides a snapshot from June 2010 on a selection of test and demographic variables.

(Table 14.1 here)

Section Summary: AphasiaBank uses a uniform protocol for the collection of discourse and demographic data, as well as testing of participants. The AphasiaBank protocol and demographics collection methods are available on the web at http://www.talkbank.org/AphasiaBank.

14.5 Transcribing

All discourse samples are transcribed in the CHAT format (MacWhinney 2000). CHAT is a transcription format 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. The CHAT transcription format is designed to operate closely with a set of programs called CLAN, which is also described in MacWhinney (2000). These programs, along with electronic versions of the manual, can also be downloaded from the AphasiaBank website at http://talkbank.org/AphasiaBank.

The CLAN programs permit the analysis of a wide range of linguistic and discourse structures. Transcription in CHAT is facilitated by a method called Walker Controller, which allows the transcriber to continually replay the original audio record. This method is built into the CLAN program and the editing of transcripts relies on the CLAN editor facility. One direct result of this process is that each utterance is then linked to a specific region of the audio or video record. This linkage can be useful for verification of transcription accuracy and for later phonological, gestural, or conversational analysis. A transcription training manual was prepared specifically for AphasiaBank purposes and posted at the website. Following the guidelines set by Berndt et al. (2000), utterances are segmented based on the following hierarchy of indices: syntax, intonation, pause, and semantics. Many students and research assistants in our facility and others have been trained to transcribe reliably and every transcript is reviewed by at least two transcribers for accuracy. For the aphasia transcripts, one of those reviewers is always speech-language pathologist.

The following CHAT fragment from the elman07a file in AphasiaBank displays some of the basic CHAT coding conventions for marking linguistic behaviors such as word repetitions ([/]), fillers (&), and gestures (&=).

```
(1)   *INV: what kinds of things have you done to try to get better
            since your stroke ?
      *PAR: &uh &=shrugs hell I don't know .
      *PAR: I suppose &uh everything [/] &uh everything better all the
            time .
```

At the end of each line, there is a round bullet symbol that contains information regarding the time value for the beginning and end of the utterance. Usually, this bullet is closed. However, if you wish to see these values, the bullet can be expanded to display the times, as shown here:

```
(2)   *INV: what kinds of things have you done to try to get better since
            your stroke ? •244832_249548•
      *PAR: &uh &=shrugs hell I don't know . •249548_257129•
      *PAR: I suppose &uh everything [/] &uh everything better all the
            time . •257129_261774•
```

Section Summary: AphasiaBank transcription relies on the CHAT data format as formalized in the CHAT manual available from http://talkbank.org/AphasiaBank.

14.6 Error Coding

Errors are coded at both the word and sentence level by speech-language pathologists. For word-level errors, we have developed a hierarchical system to capture errors in six categories: phonology, semantics, neologism, dysfluency, morphology, and formal lexical features. Within each category, errors are coded further to capture whether the error was a word or non-word, whether the target was known or unknown, whether a suffix was missing, and more. Errors that are not real words are transcribed using IPA. The error code can also indicate if the error was repeated or changed (retraced) by the speaker within the utterance by added "-rep" or "-ret", respectively, to the error code. Examples (3a-f) illustrate the six word-level error types.

(3a) Neologism, unknown target [* n:uk]:

```
*PAR: and of course she has a fancy ɹup@u [* n:uk].
```

(3b) Phonological error, real word, target known [* p:w]:

```
*PAR: and she went to the [/] &uh the mall [: ball] [* p:w] .
```

(3c) Phonological error, non-word, target known [* p:n]:

```
*PAR: peanut bʌθə̩@u [: butter] [* p:n] and sɛlɪ@u [: jelly] [* p:n]
      sæmɪtʃ@u [: sandwich] [* p:n]
```

(3d) Semantic error, related word, known target [* s:r]:

```
*PAR: and the &m mother has two daughters himself [: herself] [* s:r] .
```

(3e) Semantic error, unknown target [* s:uk]:

```
*PAR: and they get married and live everyone [* s:uk] you know .
```

(3f) Morphology error, overregularized [* m:=s]:

```
*PAR:  &uh my second third and fourth childs [* m:=s]
       were &=laughs +...
```

The reader will note that these examples included some CHAT symbols that had not yet been covered: @u for “Unicode” is appended to all IPA productions; intended (target) words, if known, are placed next to the error production as [: target]; errors are coded as [* error code] immediately following the target word, if known, or the error itself; and +... at the end of an utterance indicates trailing off.

In addition to these six categories of word-level codes, there are several utterance-level codes that are marked at the ends of utterances, as illustrated in (4a-c):

(4a) Agrammatism [+ gram]:

```
*PAR:  yeah &=finger:write May twenty fifth two thousand one
       I [/] I &s I [/] I am a stroke . [+ gram]
```

(4b) Empty speech [+ es]:

```
*PAR:  &uh I went to my &=sighs whatever &=laughs . [+ es]
```

(4c) Jargon [+ jar]:

```
*PAR:  if I could fɹeɪv@u [: x@n][* n:uk] it I guess I can
        bɹæm@u [: x@n] [* n:uk] it. [+ jar]
```

{I don't see an explanation for `[: x@n]`***; should this be included? (Just ignore, if I've overlooked it!}***

14.7 Analyses

14.7.1 CLAN

Once files have been transcribed in CHAT, users can run a wide variety of CLAN analysis programs. There are 29 CLAN 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, and so on. It is worth noting that there are several fields for demographic information to be included in the header lines of a transcript. This allows for the analysis outputs to include that information or for analyses to be conducted on particular subsets of the data, for example males versus females or just participants with Wernicke's aphasia.

14.7.2 Extensible Markup Language (XML)

The TalkBank project, which includes AphasiaBank as well as several other shared databases, has constructed Java-based tools that convert CHAT files to XML. The XML format facilitates systematic analysis, display, and searching of data over the web, but it is intended for reading by programs, not by humans. These XML files can then be reformatted back to CHAT and the initial and final versions compared to guarantee the accuracy of the roundtrip. Only when the roundtrip runs without differences can we accept the data into TalkBank. The process of converting the database to XML was completed in 2004, after nearly three years of work. An important outcome of this conversion has been the full systematization of the coding system and an increase in the consistency of the database. In addition, we were able to convert a wide range of discrepant font and character encoding systems to a consistent Unicode format. This was particularly important for Asian languages that use non-Roman characters, but it was also useful for special Roman characters with diacritics in languages such as French, German, and Spanish.

14.7.3 GEM

When transcribing language samples that include various tasks, headers marked by @G or "gem headers" are used to mark the beginning of a new task. In AphasiaBank, for instance, some of the headers are @G: Umbrella (for the refused umbrella picture stimulus) and @G: Cinderella (for the Cinderella story telling). For example, if you wanted to look at the Cinderella portion of a transcript only and you wanted the participants' lines only, you would use this command:

```
      gem +sCinderella +t*PAR +n +d1 +f *.cha.
```

The result would be new gem files in legal CHAT format with file names, line numbers, and identification codes for each original CHAT file in the folder. If you wanted to do this on multiple folders, you simply need to add +re to the command line.

14.7.4 Lexical and morphological coding

CLAN has a subprogram called MOR that applies part-of-speech taggers for English, Spanish, German, French, Italian, Japanese, Cantonese, and Mandarin. The results of these taggers are then disambiguated using the statistical disambiguator called POST (Parisse & Le Normand 2000) that uses the context before and after the word to assign part-of-speech to ambiguous cases. The transcript then appears with a new tier, %mor, under each speaker tier that gives the lexical and morphological coding for each word on the main speaker tier. These morphological codes can then be used to automatically compute indices such as DSS (Developmental Sentence Score, Lee 1966), IPSyn (Index of Productive Syntax, Scarborough 1990), and a simple version of LARSP (Language, Assessment, Remediation, and Screening Procedure, Crystal et al. 1976).

The following example transcript shows a block of conversation that has been automatically supplemented with a %mor line in which each word of the main line is given a full morphological analysis.

```
(5)   *INV: can you tell me more &=ges:more about it ?
      %mor: aux|can pro|you v|tell pro|me adv|more prep|about pro|it ?
      *PAR: well my stroke started on [//] &uh &w (.) one night and I did not
            think it was too bad .
      %mor: co|well pro:poss:det|my n|stroke v|start-PAST pro:indef|one
            n|night conj:coo|and pro|I aux|do&PAST neg|not v|think pro|it
            v:cop|be&PAST&13S adv:int|too adj|bad .
      *PAR: well it started at noon actually .
            %mor: co|well pro|it v|start-PAST prep|at n|noon
            adv:adj|actual-LY .
      *PAR: and then I went out with my friends and &uh they were concerned
            because I was driving erratically .
      %mor: conj:coo|and adv:tem|then pro|I v|go&PAST adv:loc|out prep|with
            pro:poss:det|my n|friend-PL conj:coo|and pro|they aux|be&PAST
            part|concern-PERF conj:sub|because pro|I aux|be&PAST&13S
            part|drive-PROG adv:adj|erratic-AL-LY .
```

In this sample, the main speaker tier is followed by the %mor tier. On the main speaker tier, [/] indicates repetition, [//] indicates revision, (.) indicates a short pause, & is used before fillers and word fragments, and &= is used before gestures. 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., &PAST for irregular past, -PL for regular plural, -PROG for progressive).

In this example, the %mor line was created automatically through these three computer commands:

```
mor *.cha
post *.cha
check *.cha
```

The first command runs the MOR grammar for English on the basic transcript file. This grammar can be downloaded from http://childes.psy.cmu.edu/morgrams/. The second command automatically disambiguates alternative readings inserted by MOR. The third command checks to make sure that the output is complete and syntactically accurate. For more information on the development and inner workings of MOR, POST, and CHECK, the reader is encouraged to read MacWhinney (2008).

14.7.5 Lexical diversity analysis

A glossary of CLAN commands for some basic types of analyses was developed and posted at the AphasiaBank website. The commands included in the glossary were intended to serve as a template to allow aphasia researchers to explore the wide variety of analyses that are possible. For example, to analyze lexical diversity, a researcher could use the command:

```
        vocd +r6 +t*PAR *.cha
```

to calculate VOCD (VOCabulary Diversity) in the participants' utterances for all of the CHAT files within a folder. The +r6 part of the command is used to exclude retracings (revisions) from the calculation. VOCD was developed by Malvern, Richards, Chipere, and Purán (2004) as a replacement for the type/toke ratio (TTR) measure, which fails to correct for sample size. The TTR is a simple ratio of the types of words used by a speaker in a transcript over the total number of words in the transcript. For example, if the speaker uses 30 different words and produces a total output of 120 words, then the TTR is 30/120 or .25. However, small transcripts often have inaccurately high TTR ratios, simply because they are not big enough to allow for word repetitions. VOCD corrects this problem statistically for all but the smallest samples (for details, see the CLAN manual available online at http://childes.psy.cmu.edu/). One can compute VOCD either from the main speaker line or the %mor line in the CHAT transcript. However, the goal of both TTR and

VOCD is to measure lexical diversity. For such analyses, it may not be appropriate to treat variant inflected forms or derivations of the same base (e.g., *marry, remarry,* and *married*) as different. To avoid this problem, one can compute VOCD from the %mor line using this command to control the filtering of affixes:

```
        vocd +t%mor –t* +s"*|*-%%" +s"*|*&%%" *.cha.
```

It may also be necessary to exclude other unwanted items such as neologisms or unintelligible utterances, which can be done by adding those exclusions to the CLAN command.

Fergadiotis et al. (2010) used the VOCD command to determine if productive vocabulary differs across discourse types in non-aphasic young adults (20-29 years old, n=43) and older adults (70-79 years old, n=43). (These participants are part of the non-aphasic corpus in the AphasiaBank database.) Results indicated that lexical diversity was influenced by discourse type and age. For both groups, the lexical diversity hierarchy was the same, with procedural discourse yielding the least lexical diversity, personal recounts the greatest, and single picture description and story telling falling in between. Age was a factor for the procedural discourse and personal recounts, with older adults producing significantly greater lexical diversity than the younger adults. It would be interesting to conduct these types of analyses on the discourse samples from participants with aphasia to add to our understanding of discourse and the influence of the various methods used for its evaluation and treatment.

14.7.6 MORtable

A relatively new CLAN analysis program, MORTABLE, was developed to create a table of **parts of speech** and **bound morphemes**. The command

```
mortable +t*PAR +u *.cha
```

generates a file that can be opened directly as an Excel spreadsheet. The columns of this spreadsheet provide the following information:

- Identifying information from the header lines in the CHAT transcript (e.g., participant ID, gender, type of aphasia);
- Parts of speech frequency information for wh-words, adjectives, adverbs, auxiliaries, complements, conjunctions, determiners, infinitives, modals, nouns, negations, prepositions, pronouns, possessive pronouns, reflexive pronouns, quantifiers, and verbs; and
- Bound morpheme frequency information for third person singular irregular, past irregular, third person singular regular, past regular, comparative, superlative, irregular plural, regular plural, possessive, past participle irregular, past participle regular, present participle.

14.7.7 Lexical frequency analysis

Lexical frequency analyses have been conducted to examine the Cinderella story telling lexicons in participants with and without aphasia (MacWhinney et al. 2010). The following FREQ command was used to compute the frequencies of word form occurrences on the %mor line of Cinderella gem files:

```
freq +t%mor -t* +s@r-*,o-% +u +o +fS * .gem.cex
```

This command has eight segments, the meanings of which are:

```
freq           activates the FREQ command
+t%mor         includes information from the %mor line
-t*            excludes information from the main speaker line
+s@r-*,o-%     finds all stems and ignores all other markers
+u             merges all specified files together
+o             sorts output by descending frequency
+fS            sends output to a file
*gem.cex       runs the command on all files with that extension
```

The resulting CLAN output lists the frequencies of each word used in the participants' stories. When the story transcripts include errors for which the intended target is known (e.g., *sippers* for *slippers*), the analysis will be based on tallies of the intended word (*slippers*). One can also decide whether to exclude various tokens that are counted as words, such as neologisms, unintelligible words, onomatopoeia, and letters (of the alphabet). If, for example, one wanted to exclude neologisms and unintelligible words, one would include -s@"|-neo,|-unk" in the command line.

The results of this analysis showed that non-aphasic speakers (n=25) generated 839 different word types and a cumulative total of 13,309 words; participants with aphasia (n=24) generated 526 word types and a cumulative 5,330 tokens. Examination of the word totals showed that, for each group, roughly 1/3 of the words occurred only once, another 1/3 occurred from two to four times, with the remaining 1/3 occurring five times or more. Although this wide range of lexical diversity is of interest in itself, the core ideas of the Cinderella story appear to be captured in the 306 words that occurred at least five times in the non-aphasic sample. These words included nouns, verbs, adjectives, and adverbs.

To create a target lexicon, we narrowed our focus to nouns and verbs. To search for nouns only, the following command was used:

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

The primary modification to this command from the previously explained command is the addition of |-n:*,|-n which finds all nouns, including proper nouns, compound nouns, and nouns with prefixes, still collapsing them across stems. To search for verbs, the command used these +s switches to search for verbs, auxiliaries and participles:

```
+s@r-*,|-v*,o-% +s@r-*,|-aux*,o-% +s@r-*,|-part*,o-%
```

The results showed that speakers with aphasia produced only 2/3 as many different word types as did the non-aphasic speakers, with less than half the number of tokens. Non-aphasic speakers used 80 nouns and 71 verbs at least five times. In comparison, speakers with aphasia used 34 nouns and 36 verbs five times or more, reflecting the far more restrictive lexical diversity imposed by aphasia. Nevertheless, 76% of the nouns used by the aphasic speakers also appeared in the non-aphasic lexicon.

The 10 most frequently occurring nouns in both the non-aphasic and the aphasic samples had six words in common: *Cinderella*, *ball*, *prince*, *slipper*, *mother/stepmother*, and *sister/stepsister*. The four other most frequent nouns in the aphasia stories were *man*, *shoe*, *girl*, and *home*, which are not as tightly and specifically linked to the Cinderella story as are the four other words from the top 10 nouns in the non-aphasia stories, which were *dress*, *fairy, daughter/stepdaughter,* and *godmother*.

There were eight verbs in common among the "top 10" of the aphasia and non-aphasia story samples, and all 33 verbs used by speakers with aphasia were found in the non- aphasic lexicon. Gordon (2008) tracked the usage of 11 light verbs (*be, have, come, go,* *give, take, make, do, get, move*, and *put*). All of these, with the exception of *move* and *get*, occurred in the aphasic sample, whereas only six of them appeared in the non-aphasic lexicon. The fact that the non-aphasic verb lexicon (71 verbs) was more than twice as large as the sample provided by speakers with aphasia (33 verbs) supports the argument that speakers with aphasia are in general more reliant on light verbs, showing more limited diversity for verbs. It is important to note that these analyses were conducted on 25 speakers with various types of aphasia, but with a greater representation of speakers with anomia and conduction aphasia and only a few individuals with Broca's aphasia.

To illustrate the application of these findings on an individual basis, MacWhinney et al. (2010) examined Cinderella lexicons for two speakers with different aphasia types and severities. Speaker 1 has severe Wernicke's aphasia (WAB AQ = 28.2) as a result of a stroke. He was 4 years post-onset of his aphasia, and had received both individual and group therapy since that time. Speaker 2, although scoring above the WAB cut-off for aphasia, has persistent mild word-finding problems. He displays many hesitancies and false starts of the type that characterize speakers with anomia. One of the authors (ALH) has followed this individual since his stroke approximately 10 years ago. Throughout the decade, he has received extensive individual and group treatment, and has made significant progress in rehabilitation. These two fluent speakers represent extremes of the aphasia severity scale, and not only should contrast with each other in their Cinderella narratives, but Speaker 2 should also more closely approximate the non-aphasic speech sample than he does the aphasic sample overall. If there is merit in comparing such individuals to non-aphasic speakers, then their similarities and differences from the normal lexicon should become apparent.

Results revealed that Speaker 1's total speech output for the Cinderella story was 107 words, representing 59 different word types. Accordingly, his TTR (.55) is considerably higher than the aphasic mean TTR (41). In fact, Speaker 1 used 42 words of his 107-word narration only once. Largely, this reflects his unfocused and neologistic output. However, as mentioned earlier, the TTR measure fails to correct for sample size. Using the version of VOCD built into CLAN, we found that his lexical diversity score was 45.95. However, seven of his "words" were in fact neologisms for which no clear referent could be identified. Only three nouns (*Cinderella, home, party*) and three verbs (*go, have, think*) from his sample also appeared in the non-aphasic lexicon.

In contrast, Speaker 2's narrative was both longer and much more clearly related to the lexicon of the non-aphasic speakers. It included 96 word types and 263 tokens, with a resultant TTR of .36 and lexical density of 31.11, almost precisely the non-aphasic mean for TTR (.35) and lexical density. Even though his narrative was relatively brief, it provided a substantially correct summary of the Cinderella story. (It is interesting to note that it also contained some words that were not in the non-aphasic lexicon at all, but were used appropriately. These included *lowly, envious*, and *smitten*.)

This research demonstrates that many of the methods for studying lexical patterns from the language acquisition research tradition can be applied directly to the study of lexical usage in participants with aphasia. The Cinderella story, for example, has frequently been used in aphasia research (Faroqi-Shah & Thompson 2007; Rochon et al. 2000; Stark & Viola 2007; Thompson et al. 1997). Both Rochon et al. and Thompson et al. have developed general systems for scoring narrative productions that have been applied to the Cinderella transcripts of individuals with aphasia. However, a surprising oversight in past research has been the lack of a non-aphasic standard for comparison. Without a baseline for how non-aphasic speakers narrate Cinderella, it is difficult to understand how measures of severity relate to normal expectations, and to evaluate the extent to which aphasic speakers can recover function. Furthermore, the various analyses of production in the Cinderella task have focused primarily on the construction of measures of morphosyntactic control. These measures include a wide diversity of counts of grammatical structures, inflectional processes, and sentence patterns. However, with the exception of a recent analysis by Gordon (2008), there has been relatively little attention to the analysis of the use of specific lexical items that play a role within the story of Cinderella. Hopefully, awareness of the tools described here can stimulate increased attention to patterns of lexical frequency, lexicon development, and lexical diversity in aphasia.

14.7.8 COMBO

This CLAN command can be used to search for a connected string of words. For example, to examine the use of *once upon a time* or *happily ever after* in Cinderella stories by aphasic and non-aphasic participants, one could use the following commands:

```
combo +t*PAR +re +d1 +sonce^upon^a^time *.cha
combo +t*PAR +re +d1 +shappily^ever^after *.cha.
```

In 120 aphasia samples, *once upon a time* occurred one time; in 101 non-aphasic samples it occurred 16 times. *Happily ever after* occurred 11 times in the 120 aphasic samples and 75 times in the 101 non-aphasic samples. In some of the aphasia samples, it should be noted, the productions were not error-free. Here are examples of some of the paraphasic errors observed:

```
*PAR: they live hevry [: happily] [* n:k] ever after .
*PAR: &uh and they're &maf haffiply [: happily] [* n:k] ever after.
```

Further investigations into idioms and formulaic speech are underway and should provide insight into the relative preservation or loss of these linguistic elements in aphasic discourse.

14.7.9 Error analysis

As mentioned above, AphasiaBank transcripts include a large number of word-level and sentence-level error codes. In CLAN, the FREQ command can be used to list and count each of these errors. CLAN can produce the results in a variety of ways. For example, using the +d2 option in the CLAN command sends the output to an Excel file. For word-level errors, using the +d6 option outputs the error production, the target word (if known), and the transcript file name. To search for semantic errors, one can use the following FREQ command:

```
freq +s"[\* s*]" +t*PAR +d6  adler12a.cha.
```

The output from this command looks as follows:

```
4 [* s-ret]
2 he [: she] [* s-ret]
1 she [: he] [* s-ret]
1 guy [: woman] [* s-ret]

3 [* s]
1 floor [: ground] [* s]
1 she [: he] [* s]
1 sandwich [: bread]
```

In these sample outputs, the [* s] means the error was a semantic paraphasia; the [*s-ret] means the participant retraced (revised) the semantic paraphasia within the utterance. All other CHAT coding symbols should already be familiar to the reader.

If more information about the error is desired, one can use the +d option, which outputs the selected errors with their frequencies and the filename and transcript line number where the error occurs. It also displays the actual transcript line from the file with the error so it can be seen in context. From this CLAN output, one can triple click on the filename information line and bring up onto the computer screen the whole transcript with the relevant line highlighted for even fuller context.

MacWhinney et al. (2010) illustrated a very simple example of tracking errors in the Cinderella story using the following command to trace variant forms of production of the word *Cinderella*:

```
freq +s"Cinderella" +t*PAR +u *.gem.cex.
```

This command tracks both correct uses of Cinderella and incorrect forms with the replacement code [: Cinderella] when the intended target was Cinderella. The results included paraphasic errors such as *Cinderenella*, *Cinderlella, Cilawella, Cilawilla, Cilawillipa* and *Secerundid*.

More investigations of word-level and sentence-level errors are underway and planned. Ideas for future studies have been posted at the AphasiaBank website. Within the domain of errors, we intend to delve further into the nature of paraphasic errors and the relative advantages of common coding systems. Neologistic errors (non-word errors that are not phonologically related to a known the target word) can be examined to determine what attributes permit listeners to grasp meaning in some cases but not in others. The range of questions that can be posed to these data, using these tools, is practically limitless.

Section Summary: AphasiaBank data can be analyzed by CLAN programs for errors, morphosyntax, lexical frequencies, syntactic patterns, and discourse patterns.

14.8 Syndrome Classification

Several classifications systems have been described and used over the many decades of aphasia research (Geschwind 1979; Luria & Hutton 1977; Schuell 1974). These systems have also received extensive criticism (e.g., Caramazza 1984; Schwartz 1984; Sundet & Engvik 1985). In AphasiaBank, participants are being classified in two ways: by the WAB and by their clinician. The eight possible WAB types are: Anomic, Conduction, Transcortical Sensory, Wernicke, Broca, Isolation, Transcortical Motor, and Global. They are based on the participants' scores on subtests in the domains of Spontaneous Speech Fluency and Information Content, Auditory Comprehension, Repetition, and Naming. Clinicians use their judgment and experience to identify the aphasia type, usually resulting in one of the same eight possible types used by the WAB.

We have conducted principle components and k-means analyses on the current AphasiaBank database to examine the agreement between clinician types and WAB types, where and why the disagreements occur, and what types of analyses and variables (e.g., adding discourse and error measures to the traditional measures of fluency, comprehension, naming, and repetition) may help improve the classification. One initial finding was that there are 9 participants who perform above the WAB cutoff for any aphasia type but who are still deemed to be Anomic by their clinicians. A second finding is that patients classified as Broca's aphasics by clinicians fall into two different clusters in the statistical analysis, as found earlier by Sundet and Engvik (1985). To complete a full classificatory analysis requires as large a data set as possible in order to properly classify types such as conduction. Moreover, we believe that further work needs to be done in terms of evaluating clinician judgments, as well as their match to WAB type (Swindell et al. 1984). Overall, the construction of syndrome classification continues to be a work in progress and a long-term goal of the study.

14.9 Content Analysis

The interview data can also be subjected to content analyses for features such as attitudes and coping strategies (Pennebaker et al. 2001). An example of content analysis of the database makes use of the GEM command to extract a section of the discourse sample in which participants with aphasia (n=71) were asked about their speech (Fromm, et al., in preparation) ***{This ref. is not in the reference list. Also, can 'in preparation' be updated?}***. Specifically, at the beginning of the testing session, participants were asked, "How do you think your speech is these days?". Responses to this question were coded by two researchers, revealing that positive responses accounted for 59% of all responses, followed by average or mixed responses (18%), negative responses (17%) and unclear/jargon responses (6%). Aphasia severity was significantly associated with the nature of the response, with higher WAB AQ scores in the positive group. Aphasia type and time post-onset were not significantly associated with the nature of response. Research of this type provides insight into aphasia participants' perceptions of their condition and can inform treatment designed to help individuals with aphasia capitalize on or develop resilience.

14.10 Profiles of Recovery Processes

We have only just begun to get repeated measurements on participants to start examining changes over time. The plan is to compare quantitative discourse measures before and after the administration of different types of treatments to evaluate their impact on recovery.

14.11 Conclusion

AphasiaBank provides both a rich database and powerful analysis techniques for improving our understanding of aphasia and its treatment. As this database grows in coverage for patient types, ages, and languages, we will be able to ask increasingly powerful questions. We encourage researchers to collect new data using the standard protocol and to contribute these data to the shared database. We also encourage researchers to use the tools that are available to conduct increasingly sophisticated studies of communication in aphasia.

Bibliography

Berndt, Rita, Sarah Wayland, Elizabeth Rochon, Eleanor Saffran & Myrna Schwartz. 2000. Quantitative production analysis: A training manual for the analysis of aphasic sentence production Hove, UK: Psychology Press.

Brink, T. L., Jerome A. Yesavage, Owen Lum, Philip Heersema, Michael B. Adey & Terrence L. Rose. 1982. Screening tests for geriatric depression. Clinical Gerontologist 1.37-44.

Brown, Roger. 1973. A first language: The early stages Cambridge, MA: Harvard.

Caramazza, Alfonso. 1984. The logic of neuropsychological research and the problem of patient classification in aphasia. Brain and Language 21.9-20.

Crystal, David, Paul Fletcher & Michael Garman. 1976. The grammatical analysis of language disability London: Edward Arnold.

Faroqi-Shah, Yasmeen & Cynthia K. Thompson. 2007. Verb inflections in agrammatic aphasia: Encoding of tense features. Journal of Memory and Language 56.129-51.

Feldman, Heidi, Janine. E. Janosky, Mark S. Scher & Nancy L. Wareham. 1994. Language abilities following prematurity, periventricular brain injury, and cerbral palsy. Journal of Communication Disorders 27.71-90.

Fergadiotis, Gerasimos, Heather H. Wright & Gilson J. Capilouto. (2010). Productive vocabulary across discourse types. Clinical Aphasiology Conference.

Folstein, Marshal, Susan Folstein & Gary Fanjiang. 2002. Mini-mental State Examination Lutz, FL: Psychological Assessment Resources, Inc.

Frederiksen, Carl, Janet Donin, Timothy Koschmann & A. Myers Kelson. 2004. Investigating diagnostic problem solving in medicine through cognitive analysis of clinical discourse. Paper presented at the Society for Text and Discourse, Chicago.

Fromm, Davida, Audrey Holland, Elizabeth Armstrong, Margaret Forbes, Brian MacWhinney, Amy Risko, & Nicole Mattison. (in press) “Better But No Cigar”: Persons with Aphasia Speak about their Speech, Aphasiology.

Geschwind, Norman. 1979. Specializations of the human brain. Scientific American.7-16.

Goldman, Ricky, Roy Pea, B. Barron & Sharon Derry (eds) 2007. *Video research in the learning sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, Jean. 2008. Measuring the lexical semantics of picture description in aphasia. Aphasiology 22.839-52.

Kaplan, Edith, Harold Goodglass & Sandra Weintraub. 2001. Boston naming test. Second edition Austin, TX: Pro-Ed.

Kertész, Andrew. 2007. Western aphasia battery. San Antonio: PsychCorp.

Labov, William. 2001. Principles of linguistic change. Vol. 2: Social considerations London: Blackwells.

Lee, Laura. 1966. Developmental sentence types: A method for comparing normal and deviant syntactic development. Journal of Speech and Hearing Disorders 31.331-30.

Luria, Alexander R. & J. Thomas Hutton. 1977. A modern assessment of the basic forms of aphasia. Brain and Language 4.129-51.

MacWhinney, Brian. 2010. Computational models of child language learning. Journal of Child Language 37.477-85.

MacWhinney, Brian, Davida Fromm, Audrey Holland, Margaret Forbes & Heather H. Wright. 2010. Automated analysis of the Cinderella story. Aphasiology 24, 856-868.

MacWhinney, Brian. 2000. The CHILDES Project: Tools for Analyzing Talk. 3rd Edition Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, Brian. 2008. Enriching CHILDES for morphosyntactic analysis. Trends in corpus research: Finding structure in data, ed. by Heike Behrens, 165-98. Amsterdam: John Benjamins.

MacWhinney, Brian & Jared Leinbach. 1991. Implementations are not conceptualizations: Revising the verb learning model. Cognition 29.121-57.

MacWhinney, Brian & Johannes Wagner. 2010. Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. Gesprächsforschung 2.1-20.

Malvern, David, Brian J. Richards, Ngoni Chipere & Pilar Durán. 2004. Lexical diversity and language development New York: Palgrave Macmillan.

Nicholas, Linda & Robert Brookshire. 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 38.145-56.

Parisse, Christophe & Marie Thérèse Le Normand. 2000. Automatic disambiguation of the morphosyntax in spoken language corpora. Behavior Research Methods, Instruments, and Computers 32.468-81.

Pennebaker, Jamie W., Martha E. Francis & Roger J. Booth. 2001. Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program Mahwah, NJ: Lawrence Erlbaum Associates.

Rochon, Elizabeth, Eleanor Saffran, Rita Berndt & Myrna Schwartz. 2000. Quantitative analysis of aphasic sentence production: Further development and new data. Brain and Language 72.193-218.

Scarborough, Hollis S. 1990.Index of productive syntax.Applied Psycholinguistics 11.1-22.

Schober, Michael & Frederick Conrad. 2006. Does conversational interviewing reduce survey measurement error? Public Opinion Quarterly 61.576-602.

Schuell, Hildred. 1974. Aphasia theory and therapy Baltimore: University Park Press.

Schwartz, Myrna. 1984. What the classical aphasia categories can't do for us, and why. Brain and Language 21.3-8.

Siegler, Robert S. 2006.Microgenetic analyses of learning. Handbook of child psychology: Volume 2: Cognition, perception, and language, ed. by D. Kuhn & R.S. Siegler, 464-510. Hoboken, NJ: Wiley.

Stark, Jacqueline A. & Marta S. Viola. 2007. Cinderella, Cinderella! - Longitudinal analysis of qualitative and quantitative aspects of seven tellings of Cinderella by a Broca's aphasic. Brain and Language 103.234-35.

Sundet, Kjetil & Harald Engvik. 1985. The validity of aphasic subtypes. Scandinavian

Journal of Psychology 26.219-26.

Swindell, Carol, Audrey Holland & Davida Fromm. 1984. Classification of aphasia: WAB type versus clinical impression. Paper presented at the Clinical Aphasiology Conference, Seabrook Island, SC.

Thompson, Cynthia K., Kirrie J. Ballard, Mary E. Tait, Sandra Weintraub & M. Marsel Mesulam. 1997. Patterns of language decline in non-fluent primary progressive aphasia. Aphasiology 11.297-321.

Yip, Virginia & Stephen Matthews. 2007. The bilingual child: Early development and language contact Cambridge: Cambridge University Press.

	Aphasia n=102	Non-Aphasia n=102
Age -- mean (s.d.)	63.8 years (12.9)	60.9 years (17.0)
Gender	35 females 67 males	55 females 47 males
Handedness	88 right 8 left 5 ambidextrous 1 unknown	88 right 10 left 3 ambidextrous
Education	15.6 years (3.0)	15.1 years (2.3)
Time post-onset	6.8 years (5.9)	
Type of aphasia (by WAB)	34 Anomic 26 Broca 14 Conduction 10 Wernicke 9 not aphasic 5 Transcortical Motor 2 Global 1 Transcortical Sensory 1 unavailable	
WAB AQ score	68.73 (21.19)	

Table 1. Demographic and test data for AphasiaBank participants