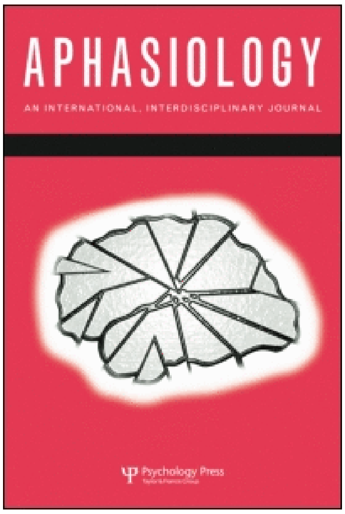


Aphasiology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713393920

Automated analysis of the Cinderella story

Brian MacWhinney[a]; Davida Fromm[a]; Audrey Holland[b]; Margaret Forbes[a]; Heather Wright[c]
[a] Carnegie Mellon University, Pittsburgh, PA, USA [b] University of Arizona, Tucson, AZ, USA [c] Arizona State University, Tempe, AZ, USA

First published on: 20 April 2010

To cite this Article MacWhinney, Brian , Fromm, Davida , Holland, Audrey , Forbes, Margaret and Wright, Heather(2010) 'Automated analysis of the Cinderella story', Aphasiology, 24: 6, 856 — 868, First published on: 20 April 2010 (iFirst)
To link to this Article: DOI: 10.1080/02687030903452632
URL: http://dx.doi.org/10.1080/02687030903452632

APHASIOLOGY, 2010, 24 (6–8), 856–868

Psychology Press
Taylor & Francis Group

Automated analysis of the Cinderella story

Brian MacWhinney and Davida Fromm
Carnegie Mellon University, Pittsburgh, PA, USA

Audrey Holland
University of Arizona, Tucson, AZ, USA

Margaret Forbes
Carnegie Mellon University, Pittsburgh, PA, USA

Heather Wright
Arizona State University, Tempe, AZ, USA

Background: AphasiaBank is a collaborative project whose goal is to develop an archival database of the discourse of individuals with aphasia. Along with databases on first language acquisition, classroom discourse, second language acquisition, and other topics, it forms a component of the general TalkBank database. It uses tools from the wider system that are further adapted to the particular goal of studying language use in aphasia.
Aims: The goal of this paper is to illustrate how TalkBank analytic tools can be applied to AphasiaBank data.
Methods &Procedures: Both aphasic (n = 24) and non-aphasic (n = 25) participants completed a 1-hour standardised videotaped data elicitation protocol. These sessions were transcribed and tagged automatically for part of speech. One component of the larger protocol was the telling of the Cinderella story. For these narratives we compared lexical diversity across the groups and computed the top 10 nouns and verbs across both groups. We then examined the profiles for two participants in greater detail.
Conclusions: Using these tools we showed that, in a story-retelling task, aphasic speakers had a marked reduction in lexical diversity and a greater use of light verbs. For example, aphasic speakers often substituted “girl” for “stepsister” and “go” for “disappear”. These findings illustrate how it is possible to use TalkBank tools to analyse AphasiaBank data.

Keywords: Lexicon; Narrative; Computer analysis.

In 2005, a group of 25 aphasiologists met to organise a proposal for a shared database on aphasia. This database was configured to operate within the framework of the larger TalkBank system that provides methods for studying a variety of language types, including child language development (childes.psy.cmu.edu), second language learning (talkbank.org/BilingBank), conversation analysis (talkbank.org/CABank),

Address correspondence to: Brian MacWhinney, Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA. E-mail: macw@cmu.edu

This project is funded by NIH_NIDCD grant R01-DC008524 (2007–2012).

http://www.psypress.com/aphasiology DOI: 10.1080/02687030903452632

phonological development (childes.psy.cmu.edu/PhonBank), legal discourse (talkbank.org/Meeting/SCOTUS), classroom discourse (talkbank.org/ClassBank), and others. The overall goal of TalkBank is to construct a shared database of multimedia data on human communication. Within the larger project, AphasiaBank focuses on the construction of a structured database that will permit the evaluation of individual differences and treatment effects in aphasia. Funding for the development of AphasiaBank was provided by NIDCD and work has been progressing on the construction of this database since 2007.

AphasiaBank collects and analyses video and audiotaped samples of the discourse of aphasic and non-aphasic participants across a wide range of tasks. One aim of AphasiaBank is to assist in the improvement of treatment for aphasia. To accomplish this, it is necessary to solidify the empirical database supporting our understanding of communication in aphasia. The eight specific aims of AphasiaBank include: protocol standardisation, database development, analysis customisation, measure development, syndrome classification, qualitative analysis, development of recovery process profiles, and evaluation of treatment effects. To advance these goals, an additional group meeting was held to formalise a shared protocol that is now available at http://talkbank.org/AphasiaBank. This protocol includes two free speech elicitation tasks, four picture description tasks, one story narrative (Cinderella), and one procedural discourse task. In addition there is a repetition test, a verb naming test (Thompson, 2010), and the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 2001). All of these tasks and tests are recorded using high-definition video and transcribed in the CHAT format (MacWhinney, 2000), with specific extensions for aphasic language. The transcripts and videos, which are password protected, can be accessed and downloaded by consortium members. Because each utterance in the transcripts is directly linked to the audio, it is possible to replay transcripts and follow along using continuous playback both over the web and locally. Participant information includes scores on the Western Aphasia Battery (WAB; Kertesz, 2007), clinical reports, and 54 demographic variables.

In this paper we focus on just one segment of this larger protocol: the telling of the Cinderella story. Within this segment we further constrain our focus to the study of patterns of lexical use in these narratives. The purpose of this paper is to provide an illustration of how one can examine substantive issues in aphasiology using this database and the CLAN programs (MacWhinney, 2000) for data analysis.

The Cinderella story has frequently been used in aphasia research (Faroqi-Shah & Thompson, 2007; Rochon, Saffran, Berndt, & Schwartz, 2000; Stark & Viola, 2007; Thompson, Ballard, Tait, Weintraub, & Mesulam, 1997). Both Berndt, Wayland, Rochon, Saffran, and Schwartz (2000) and Thompson et al. (1997) have developed general systems for scoring narrative productions that have been applied to the Cinderella transcripts of individuals with aphasia. The Cinderella story was included in the AphasiaBank protocol primarily because of its demonstrated utility, and because of its general familiarity in Western cultures. 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.

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. The study of lexical patterns in narrative has been a core topic in language acquisition studies (Malvern, Richards, Chipere, & Purán, 2004; Snow, Tabors, Nicholson, & Kurland, 1995; Tingley, Berko Gleason, & Hooshyar, 1994). Many of the methods for studying lexical patterns from this research tradition can be applied directly to the study of lexical usage in participants with aphasia. In order to take a closer look at the patterns of lexical usage in this task, we implemented a method that allowed us to contrast the patterns of lexical usage of normal participants with those of aphasic participants.

METHOD

The elicitation of the Cinderella story used the following procedure. First, participants were asked if they remembered the story of Cinderella. Then they were given a 25-page Cinderella picture book (Grimes, 2005). The text on each page of the book was covered with white duct tape to make it impossible to read. Participants paged through the book at their own pace, looking at each picture. Then the book was removed and participants were asked to tell the story of Cinderella in their own words. There was no time limit placed on their story telling. The investigator refrained from making any comments at all during the story telling. All of the productions were videotaped with audio recording that used a separate sound system.

Participants

Aphasic participants were recruited from the Adler Aphasia Center in Maywood, New Jersey, and from various venues in Tucson, Arizona, and non-aphasic participants all came from an ongoing study of normal discourse under the direction of one of the authors (HW). The aetiology for aphasia was stroke in all cases but one, which was a gunshot wound. All had been aphasic for a minimum of 6 months and a maximum of 16 years. The non-aphasic participants were screened for memory impairment (Folstein, Folstein, & Fanjiang, 2002), mood disorders, and history of stroke or other neurological conditions. The mean ages of the two groups were not significantly different. All participants had vision and hearing adequate for testing and were native speakers of standard American English. The criteria for inclusion of participants in AphasiaBank are at http://talkbank.org/AphasiaBank/inclusion.doc. Table 1 summarises demographic and other information on participant characteristics. The four participants with residual anomia tested above the cutoff on the WAB, but continued to experience and demonstrate word-finding difficulties.

Transcription

The Cinderella narratives were transcribed in the CHAT transcription format (MacWhinney, 2000). CHAT is a transcription format that has been developed over the last 30 years for use in a variety of disciplines, including first language acquisition, second language acquisition, classroom discourse, conversation analysis, etc. The CHAT transcription format is designed to operate closely with a set of programs called CLAN, which is also described in MacWhinney (2000). The CLAN programs

TABLE 1
Participant characteristics

	Non-aphasic participants (n = 25)	*Aphasic participants (n = 24)*
Age range (yrs)	23–80 (mean = 58)	30–80 (mean = 64)
Gender	16 females, 9 males	8 females, 16 males

Handedness	right = 23
	left = 1
	ambidextrous = 1
		right = 21
		left = 3

Education range (yrs)	12–20 (mean = 15)	12–25 (mean = 16)

WAB aphasia type		Anomic = 7
		Residual Anomia = 4
		Conduction = 6
		Broca = 3
		Wernicke = 3
		Transcortical Motor = 1

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 (MacWhinney, 2000) 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 second highly trained transcriber checked over the accuracy of each transcription and the two transcribers reached complete agreement on all features of the coding and transcription. Table 2 is a sample Cinderella story from participant Adler06a. This sample is a segment of a much larger transcript for the entire 1-hour interview.

The transcript includes various word-level error codes (e.g., [* wu] which indicates that the error is a real word and that the intended word is unknown) and utterance-level codes (e.g., [+ jar] for jargon) developed specifically for typical aphasic

TABLE 2
Cinderella CHAT transcript

```
@G: Cinderella
*PAR: &uh a little bit I think, yeah.
*PAR: was [//] what was the name ?
*PAR: Secerundid [: Cinderella] [* nk].
*PAR: she was &uh &b angel for legwood@n. [+ jar]
*PAR: she was &uh &f for fendle@n for someone else. [+ jar]
*PAR: the other children [/] &r &d children for her are three children or whatever . [+ es]
*PAR: with her it was very closed [* wu] walking [* wu] in generalis@n . [+ jar]
*PAR: &th &th &p pezzels@n are going for the party.
*PAR: and she was &f fen@n people [* wu] for prezzled@n (.) for the present [* wu]. [+ jar]
*PAR: the present &t (….) was s(up)posed to be &uh thirty [/] &t &uh thirty or something. [+ es]
*PAR: she &ch &er had a ranned@n from home she &ha huddled [* wu]. [+ jar]
*PAR: the &uh (..) people were +//.
*PAR: they found her letter.
*PAR: and <the pezzes@n> [//] &w the other people wed [* wu] they found her.
*PAR: found her for the prezzled@n and the calls this one so. [+ jar]
```

TABLE 3
Cinderella CHAT transcript with %mor line included

```
@G: Cinderella
*PAR: &uh a little bit I think, yeah .
%mor: det|a adj|little n|bit pro|I v|think co|yeah .
*PAR: was [//] what was the name ?
%mor: pro:wh|what v:cop|be&PAST&13S det|the n|name ?
*PAR: Secerundid [: Cinderella] [* nk] .
%mor: n:prop|Cinderella .
*PAR: she was &uh &b angel for legwood@n . [+ jar]
%mor: pro|she v:cop|be&PAST&13S n|angel prep|for neo|legwood .
*PAR: she was &uh &f for fendle@n for someone else . [+ jar]
%mor: pro|she v:cop|be&PAST&13S prep|for neo|fendle prep|for pro:indef|someone post|else .
*PAR: the other children [/] &r &d children for her are three children or whatever . [+ es]
%mor: det|the qn|other n|child&PL prep|for pro|her v:cop|be&PRES det:num|three n|child&PL
        conj:coo|or pro:wh|whatever .
*PAR: with her it was very closed [* wu] walking [* wu] in generalis@n . [+ jar]
%mor: prep|with pro|her pro|it v:cop|be&PAST&13S adv:int|very part|close-PERF
        part|walk-PROG prep|in neo|generalis .
*PAR: &th &th &p pezzels@n are going for the party .
%mor: neo|pezzels aux|be&PRES part|go-PROG prep|for det|the n|party .
*PAR: and she was &f fen@n people [* wu] for prezzled@n (.) for the present [* wu] . [+ jar]
%mor: conj:coo|and pro|she v:cop|be&PAST&13S neo|fen n|person&PL prep|for neo|prezzled
        prep|for det|the n|present .
*PAR: the present &t (.).was s(up)posed to be &uh thirty [/] &t &uh thirty or something . [+ es]
%mor: det|the n|present v:cop|be&PAST&13S adj|supposed inf|to v:cop|be det:num|thirty
        conj:coo|or pro:indef|something .
*PAR: she &ch &er had a ranned@n from home she &ha huddled [* wu] . [+ jar]
%mor: pro|she v|have&PAST det|a neo|ranned prep|from n|home pro|she v|huddle-PAST .
*PAR: the &uh (..) people were +//.
%mor: det|the n|person&PL v:cop|be&PAST +//.
*PAR: they found her letter .
%mor: pro|they v|find&PAST pro:poss:det|her n|letter .
*PAR: and <the pezzes@n> [//] &w the other people wed [* wu] they found her .
%mor: conj:coo|and det|the qn|other n|person&PL v|wed pro|they v|find&PAST pro|her .
*PAR: found her for the prezzled@n and the calls this one so . [+ jar]
%mor: v|find&PAST pro|her prep|for det|the neo|prezzled conj:coo|and det|the
        n|call-PL det|this pro:indef|one conj:subor|so .
```

language characteristics. It also includes conventional markings used by the CHAT program for repetitions ([/]), revisions ([//]), word fragments and fillers (&), replacements ([: *intended word*]), and pauses (.). The AphasiaBank website has links to a two-page sheet summarising guidelines for transcription, an error-coding document, a more detailed transcription training manual, and the complete CHAT and CLAN manuals.

The sample given in Table 2 is given again in fuller form in Table 3. The difference between Table 2 and Table 3 is that the latter includes additional material regarding part of speech tagging on the %mor line. This line gives the part of speech for each word and then provides a complete lexical analysis of the word into prefixes, stems, suffixes and clitics. It also marks whether inflectional categories are transparently analytic (as in English –ing) or fusional (as in many irregular forms), and it analyses compounds into the parts of speech of their components.

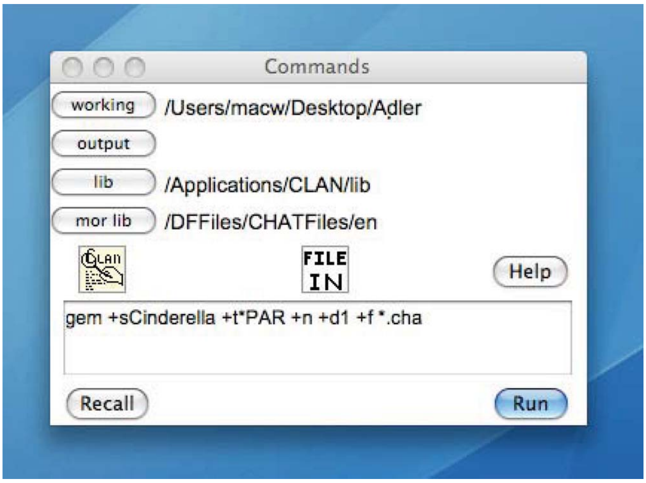
Computation of the %mor line can be done automatically, using the MOR program (Parisse & Le Normand, 2000; Sagae, Davis, Lavie, MacWhinney, & Wintner, 2007) which is included as a part of CLAN. The reader can verify that, in this example passage, all of the tags are accurate, with the exception of the last word of the last sentence that should have been tagged as an adverb. Overall, the accuracy of MOR tagging for AphasiaBank transcripts is above 98%. Although the tagger was trained on material derived from normal adult productions, it performs remarkably well at the task of tagging aphasic language.

RESULTS

To study the relative frequency of lexical items within the Cinderella story-telling task, we used a series of commands from the CLAN programs. CLAN is a single application that works on both Windows and Mac OS X (it can be downloaded from childes.psy.cmu.edu/clan). The program includes a text editor with various transcription and playback functions. There is also a commands window into which the user can type single-line commands for data analysis. The analyses presented here depend primarily on the use of these commands. In order to pull out the Cinderella story segments from the larger transcripts, we used the CLAN command called GEM. This command relies on the presence of an @G marker of the type that can be seen in the first line of Table 2 and Table 3. The specific form of the GEM command that we used was:

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

Figure 1 illustrates how this command was typed into the CLAN commands window. The result of the use of this command was a file that contained the material in Table 2. We extracted files of this type for each of our 24 aphasic and 25 non-aphasic transcripts.



working /Users/macw/Desktop/Adler
output
lib /Applications/CLAN/lib
mor lib /DFFiles/CHATFiles/en

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

Figure 1. GEM command typed into CLAN Commands window.

LEXICAL FREQUENCY ANALYSIS

To construct a lexical frequency analysis, we used the FREQ command to compute the frequencies of word form occurrences on the %mor line for each of the two folders of transcripts. The command for this was:

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

This command has eight segments. The meanings of each are as follows:

freq this calls up the FREQ command
+t%mor this includes information from the %mor line
–t* this excludes any information on the main line
+s@r-*,o-% find all stems and ignore all other markers
+u merge all specified files together
+o sort output by descending frequency
+fS send output to file
*.gem.cex run the command on all of the files with the .gem.cex extension

Table 4 shows the first lines of the output with the highest-frequency words in the stories from individuals with aphasia. This analysis is based on tallies of the intended word. Analyses of errors are beyond the scope of the current paper.

A similar analysis was computed for the non-aphasic speakers. Non-aphasic speakers generated 839 different word types and a cumulative total of 13,309 tokens; participants with aphasia generated 526 word types and a cumulative 5330 tokens. Table 5 summarises these findings and provides type token ratios, their ranges and means.

Examination of the word totals showed that, for each group, roughly 1/3 of the words occurred only once, another 1/3 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

TABLE 4
CLAN output from FREQ command

```
489 and
323 the
300 be
170 she
133 to
118 it
116 a
106 they
97 go
93 I
80 Cinderella
80 not
78 do
75 her
69 he
```

TABLE 5
TTR results

	Non-aphasic speakers (n = 25)	*Aphasic speakers (n = 24)*
Total # of different word types used	839	526
Total # of tokens	13302	5539
TTR – mean # of types	165.2	77.54
range	68–329	21–155
TTR – mean # of tokens	532.26	222.45
range	123–1347	38–705
TTR – mean	.35	.41
range	.24–.56	.17–.72

TABLE 6
The 10 most frequent nouns for the two groups

Non-aphasic speakers (n = 25)	*Aphasic speakers (n = 24)*
Cinderella	Cinderella
ball	girl
prince	ball
slipper	prince
mother, stepmother	mother, stepmother
dress	home
daughter, stepdaughter	man
fairy	slipper
godmother	shoe
sister, stepsister	sister, stepsister

nouns, verbs, adjectives, and adverbs. For purposes of this paper, we are considering only the nouns and verbs of the non-aphasic sample as constituting a target lexicon for the Cinderella data. This initial lexicon is given in the Appendix. It is the lexicon against which the stories of the participants with aphasia will be compared.

Table 5 has already alerted readers to the comparative paucity of aphasic tokens and types, and the analysis of the aphasic narratives also presents no big surprises. As a group, speakers with aphasia provided only 2/3 as many different word types as did the non-aphasic speakers, with less than half the number of tokens. As can be seen in the Appendix, 80 nouns and 71 verbs were used at least five times by non-aphasic speakers. 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% the nouns they did use also appeared in the non-aphasic lexicon.

Tables 6 and 7 present the 10 most frequently occurring nouns and verbs in the non-aphasic lexicon and the aphasic comparison. Interestingly, the most frequently occurring nouns in both the non-aphasic and the aphasic samples have six words in common. The aphasic stories included the words *man*, *shoe*, *girl*, and *home*, which are not as tightly and specifically linked to the Cinderella story, as are the words *dress*, *fairy*, *stepdaughter*, and *godmother* that appear in the non-aphasic top 10. Nevertheless, read aloud, both noun lists sound almost like an agrammatic synopsis of the Cinderella plot. It is also of interest that none of the most frequent nouns in the non-aphasic transcripts contains even a faintly abstract noun. In fact, the entire non-aphasic lexicon has only a few nouns that could possibly be construed as abstract (*love*, *life*, *course*).

TABLE 7
The 10 most frequent verbs for the two groups

Non-aphasic speakers (n = 25)	*Aphasic speakers (n = 24)*
be	be
go	go
have	do
get	have
come	get
do	say
say	know
try	find
marry, remarry	work
know	come

Verbs (see Table 7) are equally interesting. There are 7 verbs in common among the "top 10", 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 was more than twice as large as the sample provided by speakers with aphasia 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 remember that this sample of speakers with aphasia has only a few individuals with Broca aphasia and many more with anomic and conduction aphasia.

Error analysis

This analysis of the Cinderella lexicon has focused on the semantics of the words in the story. We also used CHAT codes to track neologisms and paraphasias (although the analysis of these error patterns is outside our current scope, a description of AphasiaBank error coding categories can be found at http://talkbank.org/AphasiaBank/errors.doc). However, it may be interesting to consider just a simple example of how these errors can be tracked using CLAN commands. Specifically, the following command was used 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 uses of incorrect forms with the replacement code [: Cinderella] when the intended target was *Cinderella*. The results included paraphasic errors such as: *Cinderenella, Cinderlella, Cilawella, Cilawilla and Cilawillipa* and the example in line 4 of Table 2, *Secerundid.*

Example applications

What might be the value of lexical analysis for the study of aphasic language? At present, the analysis of discourse is largely descriptive and largely dependent on features of the discourse that are of theoretical interest to the researcher. Carefully constructed lexicons of discourse samples in measures that have general use, such the Cinderella story, would make it possible to assess the severity of an individual's discourse processing deficits in a standardised way. Knowing how much and in what ways an aphasic individual's discourse performance differs from those of non-aphasic speakers on a given task could provide a real-world approach to assessment and provide guidelines and targets for treatment. For example, the simple illustration explicated here might suggest that work on developing more precise expressions for light verbs could be beneficial both in extending a linguistic repertoire, and for moving an individual closer to normal language usage. But, more generally, what would we learn from comparing a discourse sample from a speaker with aphasia to a very well-developed narrative lexicon?

To illustrate the application of these findings, we will take a closer look at the Cinderella lexicons for two speakers with aphasia. Speaker 1 has severe Wernicke's aphasia as a result of his stroke, (WAB AQ = 28.2). He is 4 years post-onset of his aphasia, and has 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 also displays many hesitancies and false starts of the type that characterise speakers with anomia. One of the researchers (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.

Following the same procedures used to gather the group data for the comparisons presented in Table 2 and 3, these speakers' individual lexicons were extracted from the larger sample. Speaker 1's total speech output was 107 words, representing 59 different word types. Accordingly, his TTR (.55) is considerably higher than the aphasic mean TTR. In fact, Speaker 1 used 42 words of his 107-word narration only once. Largely, this reflects his unfocused and neologistic output. (Table 2 includes a coded sample of his speech.) However, the TTR measure fails to correct for sample size. This problem with TTR is corrected by the VOCD command (Malvern et al., 2004). 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*) appear 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 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 words that were not in the non-aphasic lexicon at all, but were used appropriately. These included *lowly*, *envious*, and *smitten*.)

DISCUSSION

The purpose of this paper has been to introduce readers to the value of developing an archival database for aphasic language for both research and teaching purposes. This analysis illustrated the use of a few of the many analytic tools available through AphasiaBank and how they might be applied to the development of a lexicon for a narrative task that has been used frequently in aphasia research.

Eventually, the AphasiaBank database will support a much broader set of research and clinical applications. Narrative tasks of this type can be repeated across months or years to study the course of recovery from aphasia. Or we may consider the value of pre- and post-treatment samples to measure the effects of some specified treatment on lessening the impairment of aphasia. In related work, we have also developed automated methods (Sagae et al., 2007) to analyse and evaluate syntax in aphasia.

These are big questions but there are smaller, but no less interesting, questions that can be asked of the AphasiaBank database. For example, what are the attributes of neologistic errors of speakers with aphasia that permit listeners to grasp its meaning? Are they phonologic or contextual? Do they depend on shared knowledge or are they independent of it? It is not the purview of this paper to provide a laundry list of such questions but merely to suggest that the AphasiaBank database can be used to explore many issues such as these.

Manuscript received 21 July 2009
Manuscript accepted 29 October 2009
First published online 20 April 2010

REFERENCES

Berndt, R., Wayland, S., Rochon, E., Saffran, E., & Schwartz, M. (2000). *Quantitative production analysis: A training manual for the analysis of aphasic sentence production*. Hove, UK: Psychology Press.

Faroqi-Shah, Y., & Thompson, C. K. (2007). Verb inflections in agrammatic aphasia: Encoding of tense features. *Journal of Memory and Language*, *56*, 129–151.

Folstein, M., Folstein, S., & Fanjiang, G. (2002). *Mini-mental State Examination*. Lutz, FL: Psychological Assessment Resources, Inc.

Gordon, J. (2008). Measuring the lexical semantics of picture description in aphasia. *Aphasiology*, *22*, 839–852.

Grimes, N. (2005). *Walt Disney's Cinderella*. New York: Random House.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test* (2nd ed.). Austin, TX: Pro-Ed.

Kertesz, A. (2007). *Western Aphasia Battery Revised*. San Antonio, TX: Psychological Corporation.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analysing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

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

Parisse, C., & Le Normand, M. T. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, and Computers*, *32*, 468–481.

Rochon, E., Saffran, E., Berndt, R., & Schwartz, M. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language*, *72*, 193–218.

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

Snow, C. E., Tabors, P. O., Nicholson, P., & Kurland, B. (1995). SHELL: Oral language and early literacy skills in kindergarten and first-grade children. *Journal of Research in Childhood Education*, *10*, 37–48.

Stark, J. A., & Viola, M. S. (2007). Cinderella, Cinderella! Longitudinal analysis of qualitative and quantitative aspects of seven tellings of Cinderella by a Broca's aphasic. *Brain and Language*, *103*, 234–235.

Thompson, C. K. (2010). *Northwestern assessment of verbs and sentences – experimental version*. Evanston, IL: Northwestern University Press. Manuscript in preparation.

Thompson, C. K., Ballard, K. J., Tait, M. E., Weintraub, S., & Mesulam, M. (1997). Patterns of language decline in non-fluent primary progressive aphasia. *Aphasiology*, *11*, 297–321.

Tingley, E., Berko Gleason, J., & Hooshyar, N. (1994). Mothers' lexicon of internal state words in speech to children with Down syndrome and to nonhandicapped children at mealtime. *Journal of Communication Disorders*, *27*, 135–156.

APPENDIX

Cinderella lexicon of nouns and verbs for non-aphasic participants (in order of decreasing frequency)

Nouns (n = 80)

Cinderella
ball
prince
slipper
mother, stepmother
dress
daughter, stepdaughter
fairy
godmother
sister, stepsister
glass
home
girl
time
house
pumpkin
midnight
mouse
carriage
foot
father
shoe
coach
lady
animal
horse
clock
kingdom
chore
king
love
story
wife
castle
invitation
person
servant
day
palace
wand
Prince Charming
clothes
course
cat
land
magic
party
stair
thing
friend
life
man
dance
door
end
footman
princess
gown
hair
maid
night
room
dog
family
piece
scene
son
step
stroke
word
ballroom
child, stepchild
meantime
messenger
o'clock

Verbs (n = 71)		
be	think	see
go	appear, disappear, reappear	bring
have	strike	give
get	send	start
come	tell	must
do	wear	decide
say	excite	fall
try	put	pass
marry, remarry	realise	talk
know	make	want
make	let	ask
work	like	belong
fit	find	hear
find	invite	keep
see	become	push
take	help	sit
dance	meet	tear
leave	remember	happen
run	clean	end
lose	fall	happen
live	need	mean
look	treat	strike
turn	cry	