AJSLP

Research Article

Concurrent Validity and Reliability of the Core Lexicon Measure as a Measure of Word Retrieval Ability in Aphasia Narratives

Hana Kim[a] and Heather Harris Wright[a]

Purpose: General agreement exists in the literature that clinicians struggle with quantifying discourse-level performance in clinical settings. Core lexicon analysis has gained recent attention as an alternative tool that may address difficulties that clinicians face. Although previous studies have demonstrated that core lexicon measures are an efficient means of assessing discourse in persons with aphasia (PWAs), the psychometric properties of core lexicon measures have yet to be investigated. The purpose of this study was (a) to examine the concurrent validity by using microlinguistic and macrolinguistic measures and (b) to demonstrate interrater reliability without transcription by raters with minimal training.

Method: Eleven language samples collected from PWAs were used in this study. Concurrent validity was assessed by correlating performance on the core lexicon measure with microlinguistic and macrolinguistic measures. For interrater reliability, 4 raters used the core lexicon checklists to score audio-recorded discourse samples from 10 PWAs.
Results: The core lexicon measures significantly correlated with microlinguistic and macrolinguistic measures. Acceptable interrater reliability was obtained among the 4 raters.
Conclusions: Core lexicon analysis is potentially useful for measuring word retrieval impairments at the discourse level. It may also be a feasible solution because it reduces the amount of preparatory work for discourse assessment.

Discourse deficits that negatively impact daily communication for persons with aphasia (PWAs) are well known. As such, a variety of approaches for evaluating the meaningful changes in discourse for PWAs have garnered considerable attention in recent years. However, research findings that theoretically further a better understanding of how discourse deficits manifest have not resulted in clinical usability, which remains a matter of current issue. Some researchers have addressed the difficulties of discourse analysis from the clinicians' point of view (Armstrong, 2000; Bryant, Spencer, & Ferguson, 2017; Maddy, Howell, & Capilouto, 2015; Prins & Bastiaanse, 2004). Many clinicians rely on their own insights based on clinical observations when evaluating discourse ability of their patients because of difficulties in transcribing and the burden of such analyses. Yet, even in the cases when clinicians are able to collect and analyze patients' discourse samples, barriers are generally encountered that are not easily overcome, such as lack of time, limited standardized data, and no formal training programs.

To date, a variety of measures have been used to identify deficits at the discourse level in PWAs such as correct information unit (IU; Nicholas & Brookshire, 1993) and main concept (MC) analysis (Nicholas & Brookshire, 1995). A multilevel approach has more recently been suggested as a comprehensive outcome measure (e.g., Marini, Andreetta, Del Tin, & Carlomagno, 2011; Sherratt, 2007; Wright & Capilouto, 2012). Limitations of such analyses are that they require a large investment of time to train, transcribe, analyze, and interpret results and that they do not address the issue of practical application in clinical settings. Dietz and Boyle (2018) argued that there is a need for the development of ecologically valid outcome measures to evaluate discourse-level impairments. In response to their target article, other researchers presented key issues to achieve clinical use of discourse measures within clinical settings. For example, errors in segmentation of utterances and coding are likely to affect results (Kintz & Wright, 2018). Absence of

[a]Department of Communication Sciences and Disorders, East Carolina University, Greenville, NC
Correspondence to Hana Kim: only.hana.kim@gmail.com
Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Anastasia Raymer
Received March 27, 2019
Revision received June 27, 2019
Accepted August 2, 2019
https://doi.org/10.1044/2019_AJSLP-19-0063

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

criterion-referenced tools for discourse measures also hampers evidence-based practice (de Riesthal & Diehl, 2018). Finally, acceptable content validity and internal consistency should be considered for robust discourse measures (Wallace, Worrall, Rose, & Le Dorze, 2018).

Attempts to address these clinical barriers in discourse-level assessments are not new. MacWhinney, Fromm, Holland, Forbes, and Wright (2010) introduced how TalkBank tools can be applied to examine language use in discourse produced by PWAs. The authors reported that core lexicon analysis is one method to contrast patterns of lexical usage in PWAs in comparison to normal expectations. The core lexicon refers to the pivotal lexical items required to produce a semantically meaningful and coherent narrative. MacWhinney et al. used discourse samples of the Cinderella story from cognitively healthy participants ($N = 25$) in the AphasiaBank database (MacWhinney, 2000). The 10 most frequently occurring nouns and verbs were identified as core lexicon items. Then, they examined whether PWAs ($N = 24$) produced these target words to convey the Cinderella story. The PWAs demonstrated a reduced number of core lexicon items and greater use of light verbs (i.e., semantically unspecified verbs such as *be, have,* and *take*). Following similar methods, Dalton and Richardson (2015) expanded the lexical options for developing a core lexicon list. Regardless of word type, they developed a 24-core lexicon list based on language samples from 92 cognitively healthy adults from the AphasiaBank database (MacWhinney, Fromm, Forbes, & Holland, 2011). Significant differences for number of core lexicon items were found between the PWAs ($N = 92$) and control participants ($N = 166$). The researchers also used MC analysis, a measure of how accurately speakers deliver the gist of the narration, to examine the correlation between core lexicon performance and MC scores. A statistically significant correlation was found between the two measures. The researchers concluded that performance based on the core lexicon measure might reflect concept-level discourse abilities and that it may be related to PWAs' ability to construct the content of the story.

Kim, Kintz, Zelnosky, and Wright (2019) developed core lexicon lists from two narrative language samples (*Good Dog, Carl* [GDC]: Day, 1985; *Picnic*: McCully, 1984) collected from cognitively healthy participants ($N = 470$; Harris Wright & Capilouto, 2017). They considered age-related differences and word classes on usage of lexical items, which led to the development of multiple core lexicon lists based on word class (i.e., nouns, verbs, adjectives, adverbs) by age groups (20s, 30s, 40s, 50s, 60s, 70s, and 80s). Twenty-five lexical items were identified for each core lexicon list (nouns, verbs, adjectives, adverbs) among the seven age groups. Eleven PWAs were included in the study to compare their performance; percent agreement for each core lexicon list was determined. Percent agreement was calculated by dividing the number of items that each PWA produced by the total number of items (i.e., 25 items). Then, percent agreement was correlated with the overall severity of aphasia as determined by the Aphasia Quotient (AQ) from the Western Aphasia Battery–Revised (WAB-R; Kertesz, 2006). Significant correlations were found between core verbs and AQs. Core verbs also differed between persons with fluent aphasia and persons with nonfluent aphasia. In another study, Kim, Kintz, and Wright (2017) developed a 25-core function word list by using the same tasks and method. Significant correlations were found between core function word agreement and aphasia severity as measured by the WAB-R AQ.

Several researchers have reported potential advantages for using core lexicon measures for clinical practices (Dalton & Richardson, 2015; Dillow, 2013; Kim et al., 2019; MacWhinney et al., 2010). First, core lexicon lists have been developed based on cognitively healthy control participants, thus providing a norm reference for clinical populations and aiding in understanding the degree to which clinical populations deviate from typical word usage. Another advantage is that core lexicon analysis is clinician friendly. Core lexicon measures were devised to capture word retrieval ability at the discourse level by using checklists of predetermined lexical items. Instead of arduous transcription processes, potentially, clinicians can check if the predetermined lexical items are present or not while listening to language samples.

The current study serves to investigate utility of the core lexicon measure. Although previous studies have demonstrated that core lexicon measures differentiated PWAs from cognitively healthy controls (Dalton & Richardson, 2015; MacWhinney et al., 2010) and among aphasia subtypes (Dillow, 2013; Kim et al., 2019), its concurrent validity and reliability have not been investigated. In this study, we used multiple core lexicon lists derived from Kim et al. (2017, 2019; i.e., verbs, nouns, adjectives, adverbs, and function words) because of the large corpus of narrative discourse and considerations of age and word class. The purpose of the study, then, was twofold: (a) to examine the relationship among core lexicon measures and other linguistic measures (microlingusitic and macrolinguistic measures) and (b) to determine interrater reliability for the core lexicon measure using procedures and raters with minimal training. The microlinguistic measures included syntactic complexity, percentage of IUs produced, and lexical diversity using the moving average type–token ratio (MATTR) method (Covington, 2007; Covington & McFall, 2010). The macrolinguistic measures included the number of thematic units conveyed and the number of coherence units produced. Because core lexicon measures are devised to capture word retrieval ability at the discourse level and microlinguistic processes contribute to the structure of the narrative and its content (Christiansen, 1995; Ulatowska, Olness, & Williams, 2004; Wright & Capilouto, 2012), we hypothesized that performance on the core lexicon and microlinguistic measures would significantly correlate. Dalton and Richardson (2015) demonstrated that their participants who produced more core lexicon items had better MC production. Thus, we hypothesized that the core lexicon measure would significantly correlate with our macrolinguistic measures. Finally, since the core lexicon measure does not require transcription

and specific training processes, we expected clinically acceptable reliability among multiple raters.

Method

Participants

Recordings of language samples from 11 adults with aphasia (six women, five men) were used in the study, which were also included in the study of Kim et al. (2019). The participants' mean age was 61.7 (*SD* = 14.7) years, and they presented with a mean of 14.1 (*SD* = 2.7) years of education. The participants met the following inclusionary criteria: (a) native English speaker; (b) aided or unaided visual acuity as indicated by Beukelman and Mirenda's (1998) vision screening form; (c) aided or unaided hearing acuity within normal limits as measured by the ability to hear pure tones at 25 dB HL for the frequencies of 500, 1000, and 2000 Hz; (d) no reported history of psychiatric or neurodegenerative disorders; (e) a presentation of aphasia as determined by the WAB-R AQ; (f) chronic aphasia (at least 6 months post onset); and (g) left-hemisphere damage. Study participants were recruited from local support groups and university speech-language-hearing clinics. All participants provided written informed consent prior to participation. Demographic information for the PWAs can be found in Table 1.

Narrative Discourse Task

Two wordless picture books were used to collect narrative discourse samples from participants. They included *GDC* (Day, 1985) and *Picnic* (McCully, 1984). This storytelling task with visual stimuli has several advantages for eliciting language samples. First, story books following the schema of a typical Western traditional story include story elements such as setting, characters, problems, and actions. As the story proceeds, major events occur in a specific time, place, and social environment, provoking speakers' emotional response. Thus, during the task, speakers need to describe these details, thereby producing lexically diverse language samples (Fergadiotis, 2011; Fergadiotis & Wright, 2011). Moreover, pictorial support provided evokes concrete, high-imageability words (Grosjean, 1980), as well as visual imagery of the actions (Fergadiotis, 2011). Particularly, *GDC* and *Picnic* have been used as story stimuli in research (e.g., Cannizzaro & Coelho, 2013; Fergadiotis, Wright, & Capilouto, 2011; Fergadiotis, Wright, & Green, 2015; Wright & Capilouto, 2012; Wright, Capilouto, & Koutsoftas, 2013) and well investigated regarding story structures and story processing between comprehension and production (Wright, Capilouto, Srinivasan, & Fergadiotis, 2011). *GDC* is a book that illustrates the events that occur when the dog Carl is left to take care of a baby. *Picnic* is a story about a family of mice going on a picnic. For the discourse tasks, participants were presented with the book and allowed to look through it for as long as they needed to tell the whole story by themselves. In order to simulate typical clinical settings and exclude cognitive burden (e.g., memory), the books were still viewable during the course of storytelling.

Table 1. Participants with aphasia: demographic information.

Participant	Age	Gender	Education	WAB-AQ	Aphasia type
P1	65	M	18	76.3	Conduction
P3	73	M	12	85.2	Anomic
P4	84	F	12	62.6	Conduction
P5	55	M	14	57.6	Broca's
P6	66	F	14	56.3	Broca's
P7	34	F	14	90.7	Anomic
P9	38	F	14	57.7	Broca's
P10	62	F	20	61.3	Broca's
P11	72	M	12	64.9	Transcortical motor
P12	65	F	11	89.4	Anomic
P13	65	M	14	54.4	Broca's

Note. WAB-AQ = Western Aphasia Battery aphasia quotient; M = male; F = female.

Language Transcription, Measures, and Scoring

All samples were either audio- or video-recorded and then orthographically transcribed using a set of programs called *Computerized Language Analysis* (MacWhinney, 2000). In order to meet the aims of the study, microlinguistic (syntactic complexity, IUs, lexical diversity) and macrolinguistic (coherence, thematic units) analyses of the transcripts were completed. Prior to analyses, samples were segmented into communication units (c-units). A c-unit can be defined as an independent clause and includes its modifiers (Loban, 1976). An example of a c-unit is as follows:

Pre–c-unit segmented sample:

There's a family of mice that live in a house in the forest and one day they decide to pack everyone one up a large family of mice into the truck and go out for a picnic the whole family.

C-unit segmented:

1. There's a family of mice that live in a house in the forest.
2. And one day they decide to pack everyone up a large family of mice into the truck and go out for a picnic the whole family (Wright & Capilouto, 2009).

Interrater and intrarater reliability for word-by-word transcription were measured for two PWAs who were randomly selected. Interrater and intrarater transcription agreements were 91% and 93%, respectively. For c-unit segmenting, interrater agreement was 83% and intrarater agreement was 89%.

Core Lexicon Measure

The core lexicon measure has been operationally defined as a tool to quantify informativeness in discourse

production (Dalton & Richardson, 2015; Kim et al., 2019). Core lexicon production was calculated by counting how many core lexical items in the list of the respective age group were produced by each PWA. For example, an aphasia speaker who is in his 60s was evaluated by using the 60s age group's core lexicon lists. Synonyms were not counted due to the importance of producing the specific target words (e.g., Andreetta, Cantagallo, & Marini, 2012; Verhaegen & Poncelet, 2013). Again, if a PWA produced any target lexical items, they would receive 1 point. Regardless of how many times the target word may have been used by the participant, only 1 point was given. To determine the percent agreement of core lexicon production, the number of core lexicon items produced was divided by the total number of lexical items in a list; that is, if a participant produces five items, then 5/25 equals 20% agreement.

Syntactic Complexity

A complexity index (CI) was calculated to measure syntactic complexity. This index was developed by Wright and Capilouto (2012) based on previous research by Schneider, Dubé, and Hayward (2005). The index provides information on the relative syntactic complexity of a given language sample by considering clausal structure and embedding (Schneider et al., 2005). Language samples were segmented into c-units, and then CI was computed by adding the total number of independent and dependent clauses and dividing by the number of independent clauses. Interrater and intrarater agreement for calculating CI was completed for 20% of the transcribed samples. All agreements were above 90%.

IUs

An IU is operationally defined as a word that is intelligible, relevant, accurate, and informative relative to the given stimulus. IUs were determined based on previously developed guidelines (Dijkstra, Bourgeois, Allen, & Burgio, 2004; Nicholas & Brookshire, 1993). To determine the percentage of IUs produced, the number of IUs was divided by the total number of words produced and then multiplied by 100. Words included all intelligible words, regardless of their relevance, accuracy, and/or informativeness relative to the stimulus. Interrater and intrarater agreement for calculating the IUs was completed for 20% of the transcribed samples. All agreements were above 90%.

Lexical Diversity

Lexical diversity refers to a speaker's range of vocabulary (Fergadiotis & Wright, 2011). Type–token ratio (TTR) has been used in past research to estimate a speaker's lexical diversity; however, researchers have reported that TTR is sensitive to sample length and results are not reliable (McKee, Malvern, & Richards, 2000). Covington and colleagues developed an alternative measure for estimating lexical diversity, MATTR (Covington, 2007; Covington & McFall, 2010). MATTR calculates lexical diversity using a moving window to estimate TTRs for consecutive nonoverlapping segments of a language sample based on a fixed window length. Based on previous research, MATTR was calculated using a 10-word-length window within Computerized Language Analysis (Fergadiotis & Wright, 2011), and then the estimated TTRs were averaged across the sample.

Coherence

Coherence is operationally defined as the maintenance of a topic within a discourse based on the raters' impressions of the meaning of the entire verbalization with respect to the discourse topic. To analyze coherence, each c-unit was coded into a linguistic unit (i.e., noun, verb, preposition, noun phrases, verb phrases, and prepositional phrases) and then evaluated by a trained rater to determine whether it counted as a coherence unit. Coherence units included actions, locations, time, objects, people, and the positions related to the discourse topic. Prior to scoring coherence, raters completed a training protocol for scoring coherence. The protocol included language samples to practice scoring and review for accuracy in scoring. Intrarater and interrater agreement for calculating coherence was then completed for 20% of the transcribed samples selected at random. All agreements were higher than 90%.

Thematic Units

Thematic units are defined as information structurally necessary to construct informative discourse (Glosser & Deser, 1992; Marini, Boewe, Caltagirone, & Carlomagno, 2005). Thematic units included elements and actions that are informative for describing the characters and concepts (elements) and the actions in the story (actions). Guidelines for what constituted a thematic unit followed those of previous studies (Kintz, Hibbs, Henderson, Andrews, & Wright, 2018; Marini et al., 2005). To identify thematic units for the stories *GDC* and *Picnic*, cognitively healthy younger adults ($N = 3$) were asked to produce a story with a beginning, middle, and end for each. The language samples were transcribed by a trained research assistant. Elements and actions that were only produced by all three adults or that all three adults agreed on were considered to be essential thematic units. For *GDC,* 15 thematic units were identified, and for *Picnic,* 12 thematic units were identified. Reliability was calculated by dividing the number of agreements by the total number of agreements and disagreements. Both intrarater and interrater reliability were above 90%.

Rater Reliability

Of all PWAs, 10 language samples were used; one PWA was excluded because the PWA did not produce both stories. Given the different level of proficiency in discourse analysis across clinicians, raters included four research assistants with varying amounts of clinical and research experience with discourse analysis procedures. Two raters were doctoral students who had 3–4 years of clinical and research experience. The other two raters were undergraduate students in communication sciences and disorders with some experience with transcribing language samples and assisting in clinical activities (e.g., aphasia support group) as volunteers. Raters were instructed not to score synonyms but to score plurals, verb conjugations, and inflections for

the target core lexicon. Prior to scoring the participants' language samples, the raters practiced scoring once using an audio file of a language sample with checklists of the core lexicons (i.e., nouns, verbs, adjectives, adverbs, function words). Raters were instructed to check the words from the core lexicon list when they heard them in the participants' stories. In an attempt to consider typical time available for clinicians in clinical settings to complete assessments, raters were able to listen to each story no more than two times for each list. The order of scoring each list within each PWA was counterbalanced.

Results

Concurrent Validity Analyses

To address the first aim of the study, Spearman's correlation coefficients were computed for the variables of interest, by story. Spearman's correlation coefficients are considered to be an appropriate correlational analysis for data that include (a) small sample sizes ($n < 30$), (b) non-normal distributions of variables, and (c) large standard deviations (Goodwin & Leech, 2006).

For *GDC*, core nouns significantly correlated with the coherence and thematic units' measures, $r = .671$, $p < .05$ and $r = .736$, $p < .05$, respectively. Core adverbs significantly correlated with IUs and lexical diversity, $r = -.673$, $p < .05$ and $r = -.661$, $p < .05$, respectively. Core function words significantly correlated with syntactic complexity, $r = .722$, $p < .05$ (see Table 2).

For *Picnic*, core verbs significantly correlated with syntactic complexity and lexical diversity, $r = .616$, $p < .05$ and $r = .630$, $p < .05$, respectively. Core nouns significantly correlated with coherence, $r = .654$, $p < .05$; thematic units, $r = .627$, $p < .05$; syntactic complexity, $r = .652$, $p < .05$; and lexical diversity, $r = .627$, $p < .05$. Core function words also significantly correlated with coherence, $r = .778$, $p < .01$; thematic units, $r = .634$, $p < .05$; syntactic complexity, $r = .803$, $p < .01$; and lexical diversity, $r = .824$, $p < .01$. Core adjectives significantly correlated with IUs and lexical diversity, $r = .636$, $p < .05$ and $r = .701$, $p < .05$, respectively. No significant correlations were found among core adverbs and the microlinguistic and macrolinguistic measures (see Table 3).

Reliability Analyses

To determine reliability coefficients, intraclass correlation coefficients (ICCs; Nunnally & Bernstein, 1994; Shrout & Fleiss, 1979) were selected. ICC is considered a more conservative index of reliability than the Pearson product–moment correlation, which has been used as a reliability measure as well (Denegar & Ball, 1993). Following Hallgren's (2012) guidelines, absolute agreement ICC was computed with SPSS statistical software (Version 22, SPSS, Inc.). Prior to statistical analysis, ICC statistic parameters were specified. Considering our study design, the model in the current study was defined as a two-way, random ICC. Since Hallgren has suggested that it is more appropriate to use raw scores for assessing reliability rather than transformation of variables, the number of core lexicon items produced was included in statistical analysis. Standard error of measurement (*SEM*) was also calculated. *SEM* estimates the likely range of true scores (Tighe, McManus, Dewhurst, Chis, & Mucklow, 2010), indicating the amount of variation in the measurement errors (Harvill, 1991). The ICC ranged from .939 to .996 for *GDC* and from .985 to .997 for *Picnic*. The *SEM* ranged from .246 to .415 for *GDC* and from .193 to .372 for *Picnic*. Table 4 includes the ICCs and *SEM* for both stories.

Discussion

The purpose of the current study was to examine whether core lexicon measures are appropriate to use for discourse assessment, potentially in clinical settings, where economy of assessment procedures is required. With respect to concurrent validity, performance of core lexicon production correlated with both microlinguistic and macrolinguistic measures. Likely due to the different story structures of the story tasks, inconsistent findings emerged within the statistical analyses. However, the ICC for both stories ranged from strong to excellent reliability, indicating that the core lexicon lists are a reliable measure of typical word usage in discourse produced by PWAs. As mentioned previously, core lexicon studies and development of the measure are in a nascent stage. This discussion is based on results obtained with a small number of PWAs' language samples. Therefore, it should be noted that the results present preliminary validity and reliability data on the core lexicon measure.

Core Lexicon and Microlinguistic Measures

It was hypothesized that the performance of core lexicon production would significantly correlate with

Table 2. Correlation coefficients (*r*) among the core lexicon lists and linguistic measures for *Good Dog, Carl.*

Linguistic measure	Verbs	Nouns	Adjectives	Adverbs	Function words
Coherence	.275	.671*	−.059	−.024	.249
Thematic units	.193	.736*	.347	−.084	.192
Information units	.073	.354	.020	−.673*	.103
Syntactic complexity	.556	.330	−.120	−.512	.722*
Lexical diversity	.281	.299	.072	−.661*	.456

*$p < .05$.

Table 3. Correlation coefficients (*r*) among the core lexicon lists and linguistic measures for *Picnic.*

Linguistic measure	Verbs	Nouns	Adjectives	Adverbs	Function words
Coherence	.584	.654*	.594	.328	.778**
Thematic units	.532	.627*	.530	.103	.634*
Information units	.359	.444	.636*	.055	.528
Syntactic complexity	.616*	.657*	.582	.283	.803**
Lexical diversity	.630*	.627*	.701*	.360	.824**

*$p < .05$. **$p < .01$.

microlinguistic measures (IUs, syntactic complexity, and lexical diversity). We found 11 statistically significant correlations among the core lexicon measures and microlinguistic measures across the stories. In the previous literature, core lexicon production was defined as the typical usage of words at the discourse level that reflects the speakers' capacity to retrieve target words (Dalton & Richardson, 2015; Kim et al., 2019; MacWhinney et al., 2010). It has also has been suggested to be a tool to measure word retrieval deficits at the discourse level (Dalton & Richardson, 2015; Kim et al., 2019).

Significant correlations were found between core function word production and syntactic complexity for both stories; PWAs with greater core lexicons for function words also produced more syntactically complex utterances. This finding is not surprising and adds empirical evidence for the utility of using core function word lists for investigating PWAs' language ability. Function word production at the discourse level is associated with more elaborate sentence structures (Halliday & Hasan, 1976). The function core word list included conjunctions and prepositions, and these function words are considered to occur with dependent clauses when calculating CI. Also, not surprisingly, core verbs and nouns significantly correlated with syntactic complexity for *Picnic*. Dependent clauses are embedded within an utterance and include content words that are likely core verbs and nouns. Traditional measures to quantify PWAs' production of function words in discourse often require clinicians to discriminate grammatical errors, which is not practical in clinical settings. However, core function word items are identified on the list, and PWAs' scores are obtained by examining the presence or absence of function words in discourse. Once clinicians are familiar with checklists consisting of core lexical items, scoring procedures to quantify word retrieval ability in discourse may be done online.

Before moving forward to the other findings, it is important to recognize structural differences of the two stories implemented in the current study (see Supplemental Material S1). The two stories present with different story structure formats such as settings and problems (Wright & Capilouto, 2012; Wright et al., 2011). *GDC* follows a temporal story structure and includes numerous details to the story. *Picnic* may be considered a more complex story structure as it is sequentially and temporally driven. Wright et al. (2011) have previously demonstrated *Picnic* has a greater variety of story elements and older adults performed significantly better on the comprehension measure for *Picnic* than for *GDC*. It is therefore reasonable to assume that *Picnic* is a "richer" story but easy to understand for older adults. Although the relationship between comprehension of pictured stimuli and discourse output has not been readily investigated, comprehending a narrative stimulus is necessary to formulate a story from a picture book (Chapman et al., 2002). For speakers to successfully construct and deliver this story, speakers need to extract meaning from the pictured content and then integrate the information with their background knowledge (Zwaan, Langston, & Graesser, 1995; but see Wright et al., 2011). Therefore, speakers may perform differentially on discourse production tasks, depending on the extent to which they are capable of incorporating comprehension of visually presented stimuli with their own knowledge and experience. Supplemental Material S2 includes two story examples produced by a participant with aphasia.

PWAs' informativeness positively correlated with greater use of core adjectives for *Picnic*, but not for *GDC.*

Table 4. Interrater correlation coefficients and standard error of measurement (*SEM*) for *Good Dog, Carl* (*GDC*) and *Picnic*.

Discourse task	Measure	Verbs	Nouns	Adjectives	Adverbs	Function words
GDC	ICC	.984	.993	.939	.988	.996
	SEM	.372	.273	.415	.246	.394
Picnic	ICC	.985	.997	.980	.986	.997
	SEM	.325	.193	.337	.283	.372

Note. All intraclass correlation coefficients (ICCs) are positive and significant ($p < .05$). *SEM* = standard error of measurement.

For the *Picnic* story, the percent IUs conveyed increased as PWAs produced more core adjectives. This finding is of particular interest, taken together with Wright et al.'s (2011) study, as the reason for this disparity between correlation coefficients in the two narrative discourse tasks may be related to the story structure. The ease of comprehension in *Picnic* compared to *GDC* likely contributed to greater production of typical adjectives required to deliver the story. Of the 11 participants, only two produced a greater number of core adjectives in *GDC* compared to *Picnic*. It is generally accepted that completing story tasks requires a variety of cognitive processes. Moreover, processing adjectives places a greater strain on processing load (Milman, Clendenen, & Vega-Mendoza, 2014). The *Picnic* story task does not excessively challenge a speaker's processing ability and may place on them an appropriate story processing load, particularly for adjective production.

It was not surprising that the significant correlation between the core adjective list and the percent IUs was found. In an earlier study, Penn (1987) suggested that an increased use of adjectives reflects elaboration of verbal messages produced in PWAs. Sarno, Postman, Cho, and Norman (2005) also suggested that production of adjectives manifested qualitative changes in PWAs' language gain over the course of language treatment. Collectively, it was reasonable to assume that the core adjective list might be measuring elaborated descriptions in story tasks affecting the greater performance on the percent IUs.

For the *Picnic* story, several significant, positive correlations (i.e., core verbs, nouns, adjectives, function words) were found with the lexical diversity measure. As hypothesized, the core lexicon measure significantly correlated with MATTR, the lexical diversity measure. PWAs with more diverse vocabulary produced greater core lexicons. In an earlier study (Dalton & Richardson, 2015), it was hypothesized that measures consisting of a limited number of predetermined lexical items (i.e., core lexicon measures) would not be positively correlated with indices measuring varying lexical items produced due to the different approaches to measuring word retrieval ability at the discourse level. An individual who produces many synonyms may not receive core lexicon "points" because the core lexicon measure only provides points for the target words; yet, synonym production can result in greater lexical diversity scores. However, the findings appear to indicate that PWAs' ability to retrieve the most typical words does not separate from their ability to produce various different words. Production of a greater number of synonyms is considered to be a manifestation of an individual's word retrieval difficulty (Andreetta et al., 2012; Verhaegen & Poncelet, 2013; but see Dalton & Richardson, 2015). Moreover, lexical diversity is involved in the process of lexical access and retrieval, which reflects knowledge or capacity of lexicons (Fergadiotis & Wright, 2011; Fergadiotis, Wright, & West, 2013). Following this conceptualization, both lexical diversity and core lexicon measures are presumably dependent on similar discourse features, such as lexical semantics.

Unlike previous evidence showing statistically strong correlations between core verbs for both stories and overall aphasia severity (Kim et al., 2019), there is a lack of consistent relationships between core verbs and microlinguistic measures across the stories. This result may have arisen because the core verb list of *GDC* has more light verbs compared to that of *Picnic* based on Gordon's (2008) definition of light verbs. As mentioned previously, the two stories employed in the current study have different story elements and structures, which likely led to the different proportion of light and heavy verbs in the core lists. In other words, for speakers to deliver the core idea of the *Picnic* story, more semantically complex verbs (i.e., heavy verbs) are required compared to when speakers tell the *GDC* story. Heavy verbs include specific meanings and are more constrained with respect to the context in which they occur. Thus, it is possible that the *Picnic* story elicits more precise, specific expression of the story by using heavy verbs, thereby capturing the richness and complexity of PWAs' verbal output (lexical diversity, syntactic complexity).

The few, significant negative correlation coefficients obtained from the *GDC* story for informativeness and lexical diversity do not provide concurrent validation for core adverb lists. The observed trends may be attributed to the nature of the measures used in this study. Core lexicon measures were devised to score word retrieval ability by checking the presence and absence of lexical items to reduce workloads for clinicians, which is different from other measures, particularly for IUs. In studies of preschool children, "proper" use of adverbs in utterances is believed to be predictive of narrative quality and comprehension (Barnes, Kim, & Phillips, 2014). Presumably, the methodological approach of core lexicon measures is unlikely to be suitable for quantifying adverb production in discourse. Admittedly, the key factor to drive the statistical finding still remains nebulous due to the lack of studies on PWAs' adverb production. Given the absence of statistical findings among the core adverb list and both informativeness and lexical diversity for *Picnic*, it is necessary to be cautious in using core adverb lists until additional experiments for refinement of core adverb lists can be completed. More data are needed and future investigations are warranted to understand adverb contributions in discourse analyses.

Core Lexicon and Macrolinguistic Measures

Supporting and extending previous research, several significant correlations emerged among the core lexicon measures and macrolinguistic measures. Dalton and Richardson (2015) found that core lexicon performance significantly correlated with main concept scores. They suggested that function words included in their core lexicon list were the main driver of the significant results. We were able to test this hypothesis by considering word-type lists separately. Function words significantly correlated with coherence and thematic unit measures for the *Picnic* story, but not *GDC*. In the aphasia literature, cognitive deficits (e.g., working memory, attention) have been reported as partly accounting

for impaired discourse coherence (e.g., Andreetta et al., 2012; Ellis, Henderson, Wright, & Rogalski, 2016; Rogalski, Altmann, Plummer-D'Amato, Behrman, & Marsiske, 2010) and reduced function word production (e.g., Kolk & Heeschen, 1996; Salis & Edwards, 2004). Possibly then for the current study, the cognitive demands for conveying the *Picnic* story affected core function word production, maintenance of discourse coherence, and conveying the thematic units, resulting in positive relationships among the measures.

We also found core nouns significantly correlated with the macrolinguistic measures for both stories. As PWAs produced more core nouns, discourse coherence also increased. These findings add to and extend previous research findings (Dalton & Richardson, 2015). Presumably, nouns play a critical role in delivery of the overall message and thematic unity, thereby conveying substantive information about the story. Moreover, it is likely that these findings were driven by collinearity among different levels of linguistic processing. For a speaker to generate a coherent discourse in response to a topic, accurate information at linguistic levels and a logical construction of propositions are required.

Rater Reliability of Core Lexicon Measure

Absolute agreement ICC was evaluated on scores (the number of core lexicon items produced) to investigate interrater reliability coefficients. In the core lexicon literature, high reliability of the core lexicon measure has been assumed based on the nature of the measure (e.g., nontranscription), although it has not been statistically investigated. Not surprisingly, results demonstrated that the core lexicon measure is a reliable method to use for scoring narrative discourse. Following Shrout and Fleiss's (1979) guidelines, the following ICCs are considered strong reliability (ICC = .705) and excellent reliability (ICC = .970). For the current study, all ICCs were greater than .705. Moreover, considering that two of the four raters had very limited clinical experience and only received a brief, one-time training session prior to scoring, findings suggest that the core lexicon measure would be a viable option to reconcile ecological validity with clinical usability. However, the *SEM* values for some of the variables were higher than expected. Large *SEM* values are not ideal for an assessment because this would indicate a large measurement error. A practical point to note is that, in the usual calculation of *SEM*, the standard deviation of the PWAs' scores is multiplied by $\sqrt{1 - \text{reliability}}$. The small number of participants presented with different levels of fluency and/or varying ranges of core lexicon performance, which resulted in large standard deviation values. As a result, further research should include a larger sample size and PWAs across the aphasia severity continuum.

Conclusions and Limitations

Results of the current study are informative, as they provide additional and empirical support for potential use of the core lexicon measure in clinical settings. We have focused on demonstrating preliminary evidence regarding the concurrent validity and interrater reliability for core lexicon analysis. Core lexicon performance by PWAs significantly correlated with microlinguistic and macrolinguistic measures, demonstrating concurrent validity for the measure. A critical methodological implication is that core lexicon analysis holds promise as a reliable measure of narrative discourse performance in PWAs, demonstrating clinically acceptable interrater reliability with a minimal training. Through this study, we have moved one step forward in showing usability of discourse measures, demonstrating validity and reliability with the use of core lexicon measures in a laboratory setting. Next steps should consider applying this measure in practice with clinicians with varying experience of discourse analysis to examine clinical feasibility.

Several clinical and methodological implications, as well as limitations of the study, need to be considered in future investigations. First, the difference in correlation results across stories might result from inherent properties of each of the stories. Such differences highlight the importance in selecting a discourse elicitation task. Although the *Picnic* story seems to provide more robust discourse and have greater potential for diagnostic purposes based on the results of the current study, the question of which story is best for core lexicon measures is still left unanswered. At the same time, other factors should be considered as well, including the small *N* (i.e., 11), the range of aphasia types included, and the types of verbs included in each story's core verb list. As suggested by Gordon (2008), persons with fluent and nonfluent aphasia produce a different proportion of light and heavy verb usage in connected speech. Our previous study also demonstrated that participants with fluent aphasia produced significantly more core verbs than participants with nonfluent aphasia (Kim et al., 2019). Collectively, further efforts are necessary, especially for core verb lists, to illustrate the utility of the measure. It may be that less variability within the aphasia group provides a more accurate, clearer understanding of the interrelationship among core verb retrieval and microlinguistic and macrolinguistic processing.

Although these findings offer preliminary evidence for some core lexicon lists reflecting linguistic processes across different levels of discourse production, other core lexicon lists (i.e., adjectives and adverbs) provided equivocal findings for reasons that are unclear. Sarno et al. (2005) found that production of modifiers manifested qualitative changes in PWAs' language usage over the course of language treatment, which does not fully account for the current findings. In the field of linguistics, it has been suggested that adverbs serve as an integral device to measure lexical variation (e.g., Lu, 2012) and language proficiency (e.g., Grant & Ginther, 2000), yet it should be noted that these findings are based on studies regarding second language learning. Additionally, other subcategorizations of adverbs (e.g., locative adverbs, prepositional adverbs, and quasinominal adverbs) have been considered to play a distinct role or characteristic in utterances (Gilquin, 2007; Pérez-Paredes, Hernández, &

Aguado-Jiménez, 2011). Therefore, research pertaining to PWAs' modifier production and their performance by different categories of adverbs should be considered.

Finally, in discourse studies, the same discourse feature is assessed by different measures having dissimilar methodological foundations (Linnik, Bastiaanse, & Höhle, 2016). Though the core lexicon measure was designed to provide information about the typicality of language use, it conceptually can be considered to index microlinguistic levels of language ability. Despite the deliberate choices of linguistic measures employed in the current study, future studies should consider other linguistic measures to substantiate validity of the core lexicon measure and provide additional, strong evidence of the scores.

Acknowledgments

This research was partially supported by National Institute on Aging Grant R01AG029476, awarded to Heather Harris Wright. We are especially grateful to the study participants. We also thank the volunteers in the Aging and Adult Language Disorders Lab at East Carolina University for assistance with transcription and language analyses.

References

Andreetta, S., Cantagallo, A., & Marini, A. (2012). Narrative discourse in anomic aphasia. *Neuropsychologia, 50,* 1787–1793.

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology, 14*(9), 875–892.

Barnes, A. E., Kim, Y.-S., & Phillips, B. M. (2014). The relations of proper character introduction to narrative quality and listening comprehension for young children from high poverty schools. *Reading and Writing, 27*(7), 1189–1205.

Beukelman, D. R., & Mirenda, P. (1998). *Augmentative and alternative communication: Management of severe communication disorders in children and adults*. Baltimore, MD: Brookes.

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31,* 1105–1126.

Cannizzaro, M. S., & Coelho, C. A. (2013). Analysis of narrative discourse structure as an ecologically relevant measure of executive function in adults. *Journal of Psycholinguistic Research, 42*(6), 527–549.

Chapman, S. B., Zientz, J., Weiner, M., Rosenberg, R., Frawley, W., & Burns, M. H. (2002). Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. *Alzheimer Disease & Associated Disorders, 16*(3), 177–186.

Christiansen, J. A. (1995). Coherence violations and propositional usage in the narratives of fluent aphasics. *Brain and Language, 51*(2), 291–317.

Covington, M. A. (2007). *MATTR user manual*. Athens, GA: University of Georgia Artificial Intelligence Center.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100.

Dalton, S. G., & Richardson, J. D. (2015). Core-lexicon and main-concept production during picture-sequence description in adults without brain damage and adults with aphasia. *American Journal of Speech-Language Pathology, 24*(4), S923–S938.

Day, A. (1985). *Good dog, Carl*. New York, NY: Scholastic.

de Riesthal, M., & Diehl, S. K. (2018). Conceptual, methodological, and clinical considerations for a core outcome set for discourse. *Aphasiology, 32,* 469–471.

Denegar, C. R., & Ball, D. W. (1993). Assessing reliability and precision of measurement: An introduction to intraclass correlation and standard error of measurement. *Journal of Sport Rehabilitation, 2*(1), 35–42.

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? *Aphasiology, 32,* 459–464.

Dijkstra, K., Bourgeois, M. S., Allen, R. S., & Burgio, L. D. (2004). Conversational coherence: Discourse analysis of older adults with and without dementia. *Journal of Neurolinguistics, 17*(4), 263–283.

Dillow, E. (2013). *Narrative discourse in aphasia: Main concept and core lexicon analyses of the Cinderella story* (Master's thesis). University of South Carolina, Columbia, SC. Available from ProQuest. (UMI No. 1542701)

Ellis, C., Henderson, A., Wright, H. H., & Rogalski, Y. (2016). Global coherence during discourse production in adults: A review of the literature. *International Journal of Language & Communication Disorders, 51*(4), 359–367.

Fergadiotis, G. (2011). *Modeling lexical diversity across language sampling and estimation techniques* (Doctoral dissertation). Arizona State University, Tempe, AZ.

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414–1430.

Fergadiotis, G., Wright, H. H., & Capilouto, G. J. (2011). Productive vocabulary across discourse types. *Aphasiology, 25*(10), 1261–1278. https://doi.org/10.1080/02687038.2011.606974

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research, 58,* 840–852. https://doi.org/10.1044/2015_JSLHR-L-14-0280

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology, 22,* S397–S408. https://doi.org/10.1044/1058-0360(2013/12-0083)

Gilquin, G. (2007). To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift Für Anglistik Und Amerikanistik, 55*(3), 273–291.

Glosser, G., & Deser, T. (1992). A comparison of changes in macrolinguistic and microlinguistic aspects of discourse production in normal aging. *Journal of Gerontology, 47*(4), P266–P272.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *The Journal of Experimental Education, 74*(3), 249–266.

Gordon, J. K. (2008). Measuring the lexical semantics of picture description in aphasia. *Aphasiology, 22,* 839–852. https://doi.org/10.1080/02687030701820063

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123–145.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28,* 267–283. https://doi.org/10.3758/BF03204386

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London, United Kingdom: Longman.

Harris Wright, H., & Capilouto, G. J. (2017). *Discourse processing in healthy aging in the United States (ICPSR36634-v1)*. Ann

Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR36634.v1

Harvill, L. M. (1991). NCME instructional module: Standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33–41.

Kertesz, A. (2006). *Western Aphasia Battery–Revised (WAB-R)*. San Antonio, TX: Pearson.

Kim, H., Kintz, S., & Wright, H. H. (2017, May–June), *Function words in narrative discourse in aphasia*. Poster presented at the Clinical Aphasiology Conference, Salt Lake City, UT.

Kim, H., Kintz, S., Zelnosky, K., & Wright, H. H. (2019). Measuring word retrieval in narrative discourse: Core lexicon in aphasia. *International Journal of Language & Communication Disorders, 54*(1), 62–78.

Kintz, S., Hibbs, V., Henderson, A., Andrews, M., & Wright, H. H. (2018). Discourse-based treatment in mild traumatic brain injury. *Journal of Communication Disorders, 76,* 47–59.

Kintz, S., & Wright, H. H. (2018). Discourse measurement in aphasia research. *Aphasiology, 32*(4), 472–474.

Kolk, H., & Heeschen, C. (1996). The malleability of agrammatic symptoms: A reply to Hesketh and Bishop. *Aphasiology, 10*(1), 81–96. https://doi.org/10.1080/02687039608248399

Linnik, A., Bastiaanse, R., & Höhle, B. (2016). Discourse production in aphasia: A current review of theoretical and methodological challenges. *Aphasiology, 30*(7), 765–800.

Loban, W. (1976). *Language development: Kindergarten through grade twelve* (Report No. 18). Urbana, IL: National Council of Teachers.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190–208.

MacWhinney, B. (2000). *The CHILDE project: The database*. Hove, United Kingdom: Psychology Press.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25,* 1286–1307. https://doi.org/10.1080/02687038.2011.589893

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., & Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiology, 24*(6–8), 856–868.

Maddy, K. M., Howell, D. M., & Capilouto, G. J. (2015). Current practices regarding discourse analysis and treatment following non-aphasic brain injury: A qualitative study. *Journal of Interactional Research in Communication Disorders, 6*(2), 211.

Marini, A., Andreetta, S., del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology, 25*(11), 1372–1392.

Marini, A., Boewe, A., Caltagirone, C., & Carlomagno, S. (2005). Age-related differences in the production of textual descriptions. *Journal of Psycholinguistic Research, 34,* 439–463.

McCully, E. (1984). *Picnic*. London, United Kingdom: Harper Collins Publishers.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*(3), 323–338.

Milman, L., Clendenen, D., & Vega-Mendoza, M. (2014). Production and integrated training of adjectives in three individuals with nonfluent aphasia. *Aphasiology, 28*(10), 1198–1222.

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

Nicholas, L. E., & Brookshire, R. H. (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*(1), 145–156.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (McGraw-Hill series in psychology)* (3rd ed.). New York, NY: McGraw-Hill.

Penn, C. (1987). Compensation and language recovery in the chronic aphasic patient. *Aphasiology, 1*(3), 235–245.

Pérez-Paredes, P., Hernández, P. S., & Aguado-Jiménez, P. (2011). The use of adverbial hedges in EAP students' oral performance. *Researching Specialized Languages, 47,* 95–114.

Prins, R., & Bastiaanse, R. (2004). Review. *Aphasiology, 18*(12), 1075–1091.

Rogalski, Y., Altmann, L. J., Plummer-D'Amato, P., Behrman, A. L., & Marsiske, M. (2010). Discourse coherence and cognition after stroke: A dual task study. *Journal of Communication Disorders, 43*(3), 212–224.

Salis, C., & Edwards, S. (2004). Adaptation theory and non-fluent aphasia in English. *Aphasiology, 18*(12), 1103–1120. https://doi.org/10.1080/02687030444000552

Sarno, M. T., Postman, W. A., Cho, Y. S., & Norman, R. G. (2005). Evolution of phonemic word fluency performance in post-stroke aphasia. *Journal of Communication Disorders, 38*(2), 83–107.

Schneider, P., Dubé, R. V., & Hayward, D. (2005). *The Edmonton narrative norms instrument*. Retrieved from http://www.rehabresearch.ualberta.ca/enni/

Sherratt, S. (2007). Multi-level discourse analysis: A feasible approach. *Aphasiology, 21*(3–4), 375–393.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: An analysis of MRCP (UK) examinations. *BMC Medical Education, 10*(1), 40.

Ulatowska, H. K., Olness, G. S., & Williams, L. J. (2004). Coherence of narratives in aphasia. *Brain and Language, 91*(1), 42–43.

Verhaegen, C., & Poncelet, M. (2013). Changes in naming and semantic abilities with aging from 50 to 90 years. *Journal of the International Neuropsychological Society, 19*(2), 119–126.

Wallace, S. J., Worrall, L. E., Rose, T., & Le Dorze, G. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? A core outcome set or greater standardization of discourse measures? *Aphasiology, 32,* 479–482.

Wright, H. H., & Capilouto, G. J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology, 23*(10), 1295–1308.

Wright, H. H., & Capilouto, G. J. (2012). Considering a multilevel approach to understanding maintenance of global coherence in adults with aphasia. *Aphasiology, 26*(5), 656–672.

Wright, H. H., Capilouto, G. J., & Koutsoftas, A. (2013). Evaluating measures of global coherence ability in stories in adults. *International Journal of Language & Communication Disorders, 48*(3), 249–256.

Wright, H. H., Capilouto, G. J., Srinivasan, C., & Fergadiotis, G. (2011). Story processing ability in cognitively healthy younger and older adults. *Journal of Speech, Language, and Hearing Research, 54*(3), 911–917. https://doi.org/10.1044/1092-4388 (2010/09-0253)

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*(5), 292–297.