MEASUREMENT OF WORD RETRIEVAL IN THE DISCOURSE OF PERSONS WITH APHASIA: STANDARD CORE LEXICON ITEM DEVELOPMENT AND PSYCHOMETRIC PROPERTIES

by

Hana Kim

August, 2020

Director of Dissertation: Heather Harris Wright, PhD

Major Department: Communication Sciences and Disorders

Core lexicon measures are a quantitative measure of lexical use in discourse for persons with aphasia. It is intended to provide clinicians with a clinician-friendly means to quantify lexical use in discourse based on normal expectations of discourse production for specific discourse elicitation tasks. The overarching aims of the current study were to develop a reliable and valid outcome measure for discourse-level assessment and to elucidate psychometric properties of the measure. The current investigation presents the early stages of development and validation of core lexicon measures.

The aim of Study I was to outline procedures regarding development of core lexicon measures and to explore how well core lexicon measures can capture overall aphasia severity. Study II was to explore the possibility of the extension of core lexicon framework by developing checklists consisting of core function words to quantify function word use in discourse produced by persons with aphasia. Study III focused on demonstrating concurrent validity and inter-rater reliability of core lexicon measures in order to demonstrate potential clinical usability of the measure. Study IV was to investigate construct validity and item-level psychometric properties of core lexicon measures.

Results from this study suggest that core lexicon measures can be a viable option as a clinical tool in clinical settings where a lack of both time and resources are devoted for discourse-level assessment. This investigation demonstrates that core lexicon measure provide valid, reliable clinical information on how persons with aphasia produce narratives in response to discourse tasks. Additionally, this study represents a useful approach to extend and facilitate discourse-level assessment by demonstrating potential use of universal core lexicon measures.

MEASUREMENT OF WORD RETRIEVAL IN THE DISCOURSE OF PERSONS WITH APHASIA: STANDARD CORE LEXICON ITEM DEVELOPMENT AND PSYCHOMETRIC PROPERTIES

A Dissertation

Presented To the Faculty of the Department of the Department of

Communication Sciences and Disorders

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy in Communication Sciences and Disorders

by

Hana Kim

August, 2020

MEASUREMENT OF WORD RETRIEVAL IN THE DISCOURSE OF PERSONS WITH APHASIA: STANDARD CORE LEXICON ITEM DEVELOPMENT AND PSYCHOMETRIC PROPERTIES

by

Hana Kim

APPROVED BY:

DIRECTOR OF DISSERTATION: _____
Heather Harris Wright, PhD

COMMITTEE MEMBER: _____
Charles Ellis, Jr, PhD

COMMITTEE MEMBER: _____
Kathrin Rothermich, PhD

COMMITTEE MEMBER: _____
Alexander M. Schoemann, PhD

DEPARTMENT CHAIRPERSON: _____
Jamie L. Perry, PhD

DEAN OF THE GRADUATE SCHOOL: _____
Paul J. Gemperline, PhD

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to Dr. Heather Harris Wright, who provided me unwavering support and relentless guidance throughout my five-year journey at East Carolina University. Her depth of knowledge, insight, and brilliance have set an excellent example that I hope to follow in my career. In addition to her scholarship, Dr. Wright gave me moral support and had unceasing patience to get me through all the difficulties that I came across as an international student. I could not have successfully left ECU without her tremendous mentorship. I would like to thank Dr. Charles Ellis for his guidance and advice, allowing me to grow as a researcher. Many thanks to Dr. Kathrin Rothermich for her careful thought and advice not only on my research, but also on my career path. Many thanks should also go to Dr. Alexander Schoemann for lending me his expertise and intuition for my dissertation. My special regards should be paid to Dr. Yolanda Holt, who was on the committee for my first year project, for her generosity and willingness to help me. In addition to my committee members, I would like to extend thanks to all of the faculty and staff in the Department of Communication Sciences and Disorders whom I received a great deal of support from.

Special thanks should go to both my former advisor, Dr. Jee Eun Sung, and Dr. Hyun Sub Sim at Ewha Womans University who inspired and encouraged me to pursue my doctoral studies. I have benefited from their encouragements and practical guidance as well as various research experiences provided in my time at Ewha Womans University.

My heartfelt words of gratitude go to my friends, Saryu Sharma and Dr. Stephen Kintz who shared many hours of travel for conferences with me. I have had the pleasure to work with you and have truly enjoyed our conversations and coffee breaks. I want to thank all of our undergraduate research assistants in the Aging and Adults Language Disorders Lab who have

worked with me for the last five years. Thank you to Adam Amorese for putting up with my emotional ups and downs. You provided me room to vent on the bad days and celebrate on the good days. I am also grateful to all of members of the Korean Presbyterian Church of Greenville and the Reverence, Dr. Gun Ho Lee for having welcomed me and making me feel a sense of belonging in the community.

Lastly, but more importantly, I would like to thank my parents, Dong Hyun Kim and Sein Jung, and my sister, Haeun Kim for their unwavering support and immeasurable love for all these years. I am forever indebted to my parents for getting me to where I am today. I am grateful to my sister for always being there for me as a best friend despite the long distance between us. Above all, I thank God for letting me get through the tough times. I trust his love, guidance, and his plan for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Discourse outcome measures have been evolving over decades. Such changes have enhanced researchers' understanding of discourse impairments in individuals with acquired neurogenic communication disorders. However, it is undeniable that clinical application and usability of the theoretically established outcome measures have been overlooked. Maddy, Howell, and Capilouto (2015) examined the extent to which clinicians have used discourse analysis in language assessment. It was found that a gap between clinicians' value of discourse analysis and their actual practice exists. Specifically, the importance of discourse analysis to evaluate patients' communicative exchanges is clearly perceived by speech-language pathologists. However, external influences such as time constraints hamper the clinical use of discourse measures. Bryant, Spencer, and Ferguson's (2017) study supports the identified barriers of discourse analysis in clinical settings. In their survey, nearly half of the clinicians reported that they have never implemented discourse analysis. They responded that the processes to elicit, transcribe, and analyze discourse samples are burdensome. A trained clinician generally requires more than four times the actual length of the discourse sample just to complete the transcription process alone (Armstrong, Brady, Mackenzie, & Norrie, 2007; Boles & Bombard, 1998; Elia, Liles, Duffy, Coelho, & Belanger, 1994; Boles & Bombard, 1998). This timeframe excludes the time required for training and analysis, thus making many analyses impractical for use in clinical settings.

How a clinically acceptable discourse measure is defined has been discussed in a recent forum of *Aphasiology* (de Riesthal & Diehl, 2018; Dietz & Boyle, 2018a, 2018b; Kintz & Wright, 2018; Kurland & Stokes, 2018; Wallace, Worral, Rose, & Dorze, 2018; Whitworth,

2018). First, a non-transcription discourse measure is critical to reduce clinicians' time and effort (McNeil, Doyle, Fossett, Park, & Goda, 2001; Olness, Gyger, & Thomas, 2012; but see de Riesthal & Diehl, 2018). This permits clinicians to achieve high reliability of discourse analysis (McNeil et al., 2001). Moreover, a norm reference for clinical populations should be provided as a solid foundation for clinical decisions (de Riesthal & Diehl, 2018; Dietz & Boyle, 2018b). This could help us gain insight into the nature of patients' language profiles and to what extent they are preserved or impaired.

In studies examining core lexicon measures (a lexical-based analysis) in persons with aphasia, it has been demonstrated that many of these issues could be addressed through use of core lexicon analysis (Dalton & Richardson, 2015; Dillow, 2013; Fromm, Forbes, Holland, & MacWhinney, 2013; MacWhinney, Fromm, Holland, Forbes, & Wright, 2010). Although the studies demonstrated core lexicon measures differentiated persons with aphasia (PWA) from the cognitively healthy controls (Dalton & Richardson, 2015; MacWhinney et al., 2010) and among aphasia subtypes (Dillow, 2011; Kim et al., 2018), some variability exists across the studies. To be specific, core lexicon lists were differently configured depending on the criteria to select the lexical items and the composition of a list. MacWhinney and colleagues (2012) generated 10 core nouns and 10 verbs using the Cinderella story-retelling task. Dalton and Richardson (2015) aggregated all word classes in a core lexicon list with the inclusion criteria that required at least 50% of lexical items be produced by the controls. Dillow (2013) also used the same inclusion criterion of 50% production, and separately created core verb and noun lists. Fromm, Forbes, Holland, and MacWhinney (2013) established 10 nouns and verbs simply based on frequency of lexical item produced in cognitively healthy adults. To determine the clinical usability and the validity of score interpretations, a systematic approach to the aspects that potentially affect the

quality of the measurement should be considered. Therefore, a series of investigations were designed, and details of the investigations follows.

Study I: Do age-based, separate core lexicon lists by word class significantly correlate with aphasia severity?

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.

The purpose of this study was to apply an age-based core lexicon list for nouns, verbs, adjectives, and adverbs for the wordless picture books, Good Dog Carl (GDC; Day, 1985) and Picnic (McCully, 1984), to determine how well the lists measure language impairments in PWA. In studies creating core lexicon lists, many researchers disregarded some word types (e.g. adjectives and adverbs) (Dillow, 2013; Fromm et al., 2013; MacWhinney et al., 2010) or combined words types to create a single core lexicon list (Dalton & Richardson, 2015). However, different words types, such as nouns, verbs, adjectives, and adverbs, carry important and unique semantic information that differentiate them (Neville, 2014). Moreover, it has been suggested that production of modifiers manifested qualitative changes in language usage for PWA (Sarno, Postman, Cho, & Norman, 2005). Additionally, previous studies have never considered age differences in lexical selection. The specific aims of the current study, then, were two-fold: (1) determine the percent agreement between groups and their core lexicon and (2) examine the correlation among lexicon lists and aphasia impairment as determined by the aphasia quotient (AQ) from the Western Aphasia Battery – Revised (WAB-R; Kertesz, 2006). Lemma forms were extracted from 470 control participants who were divided into seven age

groups. Twenty-five core lexicons were identified for four word classes (nouns, verbs, adjectives, and adverbs) among the seven age groups. Then, the nouns, verbs, adjectives, and adverbs for each PWA (N =11) were compared to the core lexicon for their respective age group. Results indicated that the percentage of agreement for each word type among the age cohorts ranged between 56% and 96%. Of the four word types, core verbs significantly correlated with the WAB-AQs for both discourse tasks. A post-hoc analysis found significant differences between fluent and non-fluent aphasia for core verbs. These findings are promising, as they broaden our understanding of how meaningful the verb core lexicon is for PWA and also have clinical implications. Verb counts (i.e., using a core verb list or counting verbs produced) might be a discourse measure that is sensitive to capturing comprehensive language ability. Finally, given that core lexicon measures are relatively novel methods, it is too early to draw a conclusion that other lists do not provide clinical information.

Study II: Does function word production of PWA significantly correlate with aphasia severity?

Kim, H., Kintz, S., & Wright, H. H. Function words in narrative discourse in aphasia. *Submitted to International Journal of Language & Communication Disorders.*

A multitude of measures have been applied to investigate discourse production in persons with aphasia (PWA). However, these measures have not been widely used in clinical settings due to resource-heavy procedures. The purpose of this study was to develop a clinically acceptable measure that evaluates function word production in discourse. To identify the core function word lists, age and narrative task were considered. Two wordless picture books were used to collect narrative language samples from 470 control speakers (20s, 30s, 40s, 50s, 60s, 70s, and 80s). Core lexicon lists were developed by identifying the 25 most frequently occurring

function words for the seven age groups, the two wordless picture books (Good Dog Carl, Picnic), and also combined across both stories. Language samples from 11 PWA were used to determine the relationship among function word production and aphasia determined by the aphasia quotient (AQ) from the Western Aphasia Battery – Revised (Kertesz, 2006). Significant differences for function word production were found between the two discourse stories, but not across the age cohorts. Because age was not a factor for the function word lists, discourse samples for the PWA were compared to core function lists for each discourse story. Then, Spearman's correlations were performed to determine the relationship among function word production and aphasia severity (WAB-R AQ). The two core lexicon lists significantly correlated with overall aphasia severity. These findings have potential clinical and methodological implications. Discourse elicitation stimuli should be considered when elicit heterogeneous usage of function words. Core function word lists may facilitate assessment of language usage at the discourse level by reducing time and resources in clinical settings.

Study III: Does the core lexicon measure demonstrate clinically acceptable validity and inter-rater reliability?

Kim, H., & Wright, H. H. (2020) Concurrent validity and reliability of the core lexicon measure as a measure of word retrieval ability in aphasia narratives. *American Journal of Speech-Language Pathology*, *29*(1), 101-110.

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 (PWA), the psychometric properties of core lexicon measures have yet to be investigated. The purpose of this study was (1) to examine the concurrent validity by using micro- and macro-linguistic measures and (2) to demonstrate inter-rater reliability without transcription by raters with minimal training.

Method: Eleven language samples collected from PWA were used in this study. Concurrent validity was assessed by correlating performance on the core lexicon measure with micro- and macro-linguistic measures. For inter-rater reliability, four raters used the core lexicon checklists to score audio-recorded discourse samples from ten PWA.

Results: The core lexicon measures significantly correlated with micro- and macro-linguistic measures. Acceptable inter-rater reliability was obtained among the four 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.

Study IV: What are best practices for developing core lexicon lists for use in research?

The primary goal of this research is to explore the best practices for developing a standard core lexicon set for use in research and clinical purposes. The central issue in constructing a new language test has focused on whether or not the intended linguistic functions will be appropriately measured. In a broad sense, reliability and validity are of particular relevance to aphasia language tests because measures of reliability and validity reflect not only reliability of the test, but the test's ability to discriminate severity. Despite the evidence that the core lexicon measure may address the issue of clinical feasibility for discourse analysis in clinical settings (Dalton & Richardson, 2015; Kim et al., 2018; MacWhinney et al., 2010), the

question of clinical feasibility as a standard measure is left open. As previously mentioned, different inclusionary criteria have been applied to select lexical items and also different discourse elicitation tasks have been used across studies. For the proposed study, we will follow our previous procedures (Kim et al., 2018) and generate multiple core lexicon lists by word class (e.g., verbs, nouns, adjectives, adverbs, function words). The specific aims are as follows:

STUDY AIMS

Aim I: Establish criteria for core lexicon list development.

There is no converging evidence from previous research with respect to criteria for lexical items. Different research groups have generated core lexicon lists based on the frequency of lexical items produced by the controls (Fromm et al., 2013; Kim, et al., 2018; MacWhinney et al., 2010), whereas other research groups required that at least 50% of core lexical items be produced by the control participants to be included in the core lexicon list (Dalton & Richardson, 2015; Dillow, 2011). To address this aim, two approaches to define the lexical items are considered and will be evaluated for (a) research and clinical usability and (b) construct validity.

Aim II: Evaluate use of standard core lexicon sets.

Although the core lexicon has proven to be reliable in previous research (Dalton & Richardson, 2015; Dillow, 2013; Fromm et al., 2013; MacWhinney et al., 2010), the validity of this measure has not been evaluated. To address this aim, (a) clinical feasibility and (b) clinical application will be investigated.

Specific Aim 2.1 –To address this sub-aim, a standard core lexicon list created through Aim 1 will be applied to a large corpus of language samples elicited by different discourse elicitation tasks.

Specific Aim 2.2 –To address this sub-aim, the standard core lexicon list will be evaluated for item analysis.

References

Armstrong, L., Brady, M., Mackenzie, C., & Norrie, J. (2007). Transcription-less analysis of aphasic discourse: A clinician's dream or a possibility? *Aphasiology*, *21*(3–4), 355–374.

Boles, L., & Bombard, T. (1998). Conversational discourse analysis: Appropriate and useful sample sizes. *Aphasiology*, *12*(7–8), 547–560.

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology*, *31*(10), 1105–1126.

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*, *39*(11), 1125–1137. https://doi.org/10.1044/2015_AJSLP-14-0161

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

de Riesthal, M., & Diehl, S. K. (2018). Conceptual, methodological, and clinical considerations for a core outcome set for discourse. *Aphasiology*, *32*(4), 469–471.

Dietz, A., & Boyle, M. (2018a). Discourse measurement in aphasia: consensus and caveats. *Aphasiology*, *32*(4), 487–492.

Dietz, A., & Boyle, M. (2018b). Discourse measurement in aphasia research: have we reached the tipping point? *Aphasiology*, *32*(4), 459–464.

Dillow, E. (2013). *Narrative Discourse in Aphasia: Main Concept and Core Lexicon Analyses of the Cinderella Story*. Columbia: University of South Carolina.

Elia, D., Liles, B. Z., Duffy, R. J., Coelho, C. A., & Belanger, S. A. (1994). An investigation of sample size in conversational analysis. In *ASHA Convention, New Orleans, LA*.

Fromm, D. A., Forbes, M., Holland, A., & MacWhinney, B. (2013). PWAs and PBJs: Language for describing a simple procedure.

Kertesz, A. (2006). *Western Aphasia Battery–Revised (WAB-R) Pro-Ed*. *Austin, TX*.

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

Kurland, J., & Stokes, P. (2018). Let's talk real talk: an argument to include conversation in a D-COS for aphasia research with an acknowledgment of the challenges ahead. *Aphasiology*, *32*(4), 475–478.

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., & Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiology*, *24*(6–8), 856–868. https://doi.org/10.1080/02687030903452632

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.

McCully, E. A. (1984). *Picnic*. Harper & Row New York.

McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology*, *15*(10–11), 991–1006. https://doi.org/10.1080/02687040143000348

Olness, G. S., Gyger, J., & Thomas, K. (2012). Analysis of Narrative Functionality: Toward Evidence-based Approaches in Managed Care Settings. *Seminars in Speech and Language*, *33*, 55–67. https://doi.org/10.1055/s-0031-1301163

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.

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 standardisation of discourse measures? *Aphasiology*, *32*(4), 479–482.

Whitworth, A. (2018). The tipping point: are we nearly there yet? *Aphasiology*, *32*(4), 483–486.

CHAPTER 2

LITERATURE REVIEW

Word Retrieval Ability in Discourse by Persons with Aphasia

Persons with Aphasia (PWA) present with word retrieval difficulty (Goodglass & Wingfield, 1997). This results in breakdown in the flow of connected speech and effective communication (Herbert, Best, Hickin, Howard, & Osborne, 2003; Hickin, Herbert, Best, Howard, & Osborne, 2007). To detect PWA's word retrieval deficits, standardized and comprehensive aphasia batteries have frequently been used in clinical settings, such as the Western Aphasia Battery – Revised (WAB-R; Kertesz, 2006) and the Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 2001) (Guo, Togher & Power, 2014; Verna, Davidson & Rose, 2009). Such tests generally require PWA to name objects at the single word level. The use of these measures has many advantages, as they are simple to administer, score, and interpret the results (Herbert, Hickin, Howard, Osborne, & Best, 2008). This also leads to achieving high test-retest reliability (Fergadiotis, Kapantzoglou, Kintz, & Wright, 2018; Fergadiotis, Kellough, & Hula, 2015; Herbert et al., 2008).

However, there is evidence that such testing batteries that assess word retrieval ability at the single word level do not provide sufficient information to identify word retrieval impairments at the discourse level (e.g., Mayer & Murray, 2003; Pashek & Tompkins, 2002; Williams & Canter, 1982). Relatively few studies have demonstrated a discrepancy in word retrieval ability between single word production and discourse production. In a single case study, Manning and Warrington (1996) reported a different performance between picture naming and connected speech. The participant (KP) was categorized as anomia type's aphasia. KP demonstrated

reduced lexical retrieval more so in connected speech than in single word naming tasks. HY also exhibited better verb production than noun production. Interestingly, noun production was much better during utterance-level tasks compared to verb production. The researchers interpreted the findings as evidence that two separate routes to retrieve nouns in isolated naming and connected speech exist.

Wilshire and McCarthy (2002) attempted to prove that individuals with aphasia are differentially influenced by context using computerized *cyclic naming tasks*. Wilshire and McCarthy made picture sets which consisted of 36 line drawings of objects and animals. Six sets of stimuli consisted of semantically related pictures, and the other six sets of stimuli consisted of semantically unrelated pictures. All participants (1 non-fluent aphasia, 1 anomic aphasia) were asked to name as many of the pictures as they could. The participant with non-fluent aphasia performed better with the unrelated sets than the semantically related sets. The participant with anomic aphasia did not show this affect. They concluded that context effects can be explained by "external" factors that come into play during the access and selection of lexical items.

With a group study of contextual effects, Williams and Canter (1982) found that effects of context appeared to be different between participants with Broca's aphasia compared to participants with Wernicke's aphasia. The researchers used two picture naming tasks, one in which word pictures were presented alone, and the other in which target pictures were presented within scenes. The participants with Broca's aphasia performed better on the isolated naming task than the naming task with scenes. The participants with Wernicke's aphasia exhibited an opposite pattern. In support of these findings, Schnur and colleagues (2006) reported that non-participants with fluent aphasia are greatly influenced by context effects. They concluded that

MacKay's model provided a useful framework to explain context effects because of automatic activation.

Some researchers have considered word class relative to the differential performance of word retrieval ability between single word and discourse levels. Mayer and Murray (2003) explored word retrieval performance between confrontation naming and discourse considering the effect of word class (noun vs verb) and aphasia severity (mild vs moderate) based on the aphasia quotients from the Western Aphasia Battery (WAB; Kertesz, 1982). Participants (N = 14) completed noun and verb confrontation naming tests and described a sequential picture that included three picture frames. Conversational speech was also elicited in a brief interview about a certain topic (e.g., family, occupation, travel). Persons with mild aphasia outperformed persons with moderate aphasia across all measures. For both groups, the scores of two discourse tasks were higher than scores of confrontation naming task for both nouns and verbs. The proportion of substantive verbs discriminated the participants with mild aphasia from the participants with moderate aphasia when using sequential picture descriptions, but not in conversational speech. No difference between word classes was found.

Pashek and Tompkins (2002) found similar results in their research. They explored lexical retrieval between discourse and confrontation-naming tasks in persons with mild aphasia. All participants (N = 20) completed object and action naming tests for confrontation naming and they were requested to explain the episodes and actions in a video clip. Both control and aphasia groups produced more nouns than verbs in their narrative discourse. The PWA demonstrated more word finding difficulties for nouns than they did for verbs across all contexts. In terms of word class, PWA produced fewer nouns than the control group did, despite the finding that there was no difference found for verbs. Moreover, Pashek and Tompkins (2002) found significantly

different performance for nouns compared to verbs; participants had more difficulty retrieving nouns on the confrontation naming tasks and narrative discourse task.

Kambanaros (2010) investigated differential patterns of noun and verb performance in both confrontation naming tasks and connected speech. Twelve bilingual fluent and anomic PWA and 12 cognitively healthy adults participated in this study. Researchers administered all tasks in both first language (L1; Greek) and second language (L2; English) for PWA. In both languages, PWA performed worse on nouns in spontaneous speech compared to the controls. PWA had more difficulty in retrieving nouns in isolation than retrieving nouns in context. For PWA, verb production was worse than noun production in naming tasks, whereas verb production was better than noun production in spontaneous speech for both languages.

Law, Kong, Lai, and Lai (2015) compared word retrieval ability of nouns and verbs in 19 Chinese speakers with anomic aphasia and cognitively healthy controls. They used nouns and verbs matched in age-of-acquisition (AoA) and familiarity in four tasks (picture naming, connected speech from picture description, procedural description, and storytelling). The results showed that control participants outperformed the persons with anomic aphasia in all conditions. Control participants performed better in picture naming tasks compared to connected speech tasks; and, performed better in noun production compared to verb production. Persons with anomic aphasia performed better on retrieving nouns than verbs in the picture naming task, but not in the connected speech tasks. They also showed better performance in picture naming than connected speech for both nouns and verbs. Overall, both groups performed best when retrieving nouns in picture naming. The authors highlighted the importance of evaluating word retrieval ability at the discourse level, especially for persons with anomic aphasia. They believed that

measuring changes in word retrieval ability in discourse is excluded from the evaluation and/or treatment in many research and clinical cases.

The overall findings on differential performance in word retrieval at the single word and discourse levels (with or without considering word classes) have not been converging; however, they have implicated that word retrieval performance in PWAs' discourse is heterogeneous depending on the context and/or word class.

Core Lexicon in Aphasia

Clinical feasibility of discourse outcome measures capturing word retrieval ability has been consistently questioned by aphasiologists. As such, there has been attempts to address the issues by developing discourse analysis that requires less investment in time and effort. Core lexicon is one such analysis procedure currently in development (Dalton & Richardson, 2015; Dillow, 2013; Fromm et al., 2013; MacWhinney et al., 2010). Core lexicon refers to the pivotal lexical items required to produce a semantically meaningful and coherent narrative (MacWhinney et al., 2010). As such, it can be expected that the core lexicon produced by individuals with impaired lexical access would be reduced.

MacWhinney and colleagues (2010) introduced a core lexicon analysis for the Cinderella story by analyzing the discourse samples from 25 healthy participants and 24 PWA. They collected the discourse samples from AphasiaBank, a collaborative project whose goal is to develop a database of language samples from PWA. The researchers used the Computerized Language Analysis program (CLAN; MacWhinney, 2000) to conduct the analysis. The MOR command automatically sorted each word within the discourse samples by parts of speech with 95% accuracy. Next, the researchers used the FREQ command to recall the most frequently used

nouns and verbs for the two groups. They found that the PWA's discourse abilities were characterized with reduced lexical diversity and greater use of light verbs (i.e. frequently occurring verbs in language samples such as *be, have, come* etc…) compared to the control group. However, the core lexicon lists only included nouns and verbs, and the researchers did not consider other word classes, such as adjectives and adverbs, which may contribute to increased lexical diversity in discourse productions (Sarno, Postman, Cho, & Norman, 2005).

Dalton and Richardson (2015) reported that a 24-item core lexicon list, independent of word class (i.e., verbs, nouns, adverbs, adjectives, etc.), discriminated between healthy controls and PWA. To develop the core lexicon list, the researchers accessed the transcripts of 92 healthy controls from Aphasia Bank. They extracted all the lemmas produced within one of the discourse tasks, a sequential picture description task (Broken Window). The lemmas were extracted by using the CLAN command - FREQ, where 24 lemmas produced by 50% or more of the control participants were included within the core lexicon. To determine if core lexicon could distinguish between the two groups, the researchers examined the transcripts of 166 healthy controls and 235 PWA. The researchers found that PWA and healthy controls produced a significantly different number of core lexical items. They concluded that the core lexicon list can reflect the participants' ability to convey the gist of a narration. However, the core lexicon list developed was not effective for discriminating among different types of aphasia. Further, the age of participants was not considered during development of the core lexicon lists.

Fromm, Forbes, Holland, and MacWhinney (2013) compared the core lexicon lists for a different type of discourse – procedural discourse (how to make a peanut butter & jelly sandwich). No differences were found between healthy controls (n = 145) and PWA (n = 141). They included additional measures as well: the number of words, the number of utterances, and

utterance duration. The healthy control group produced significantly more words and utterances, as well as had longer utterance durations compared to aphasia group. Fromm and colleagues suggested these measures reflect quantitative differences among the groups. Previously, Fergadiotis and colleagues (2011) found that procedural discourse tasks yield fewer lexical items compared to other types of discourse. Results from these studies highlight the need to consider discourse type as well.

Dillow (2011) analyzed the core lexicon lists for the Cinderella story. Dillow created core verb and noun lists following MacWhinney et al. (2010)'s procedures. In contrast to earlier studies, they attempted to add an adjective core lexicon list, but they did not include it due to their criterion to establish the lexicon. The scores for core verbs, core nouns, and the entire core lexicon lists differentiated aphasia subtypes from the control group. They also analyzed how different word types of core lexicon affected the ability to divide each of aphasia subtype categories. Core verbs differed for the following groupings: adults with anomic aphasia and adults with Broca's aphasia; adults with anomic aphasia and adults with conduction aphasia; and adults with Broca's aphasia and adults with Wernicke's aphasia. For core nouns, the participants with anomic aphasia significantly differed from those with Broca's aphasia and Wernicke's aphasia; and, participants with conduction aphasia also significantly differed from those with Broca's aphasia and Wernicke's aphasia. When considering the complete lexicon, adults with Broca's aphasia differed significantly from the adults with anomic aphasia and conduction aphasia. These findings suggest that core lexicon lists separated by different word types might prove more useful than a combined single core lexicon list.

References

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*, *39*(11), 1125–1137. https://doi.org/10.1044/2015_AJSLP-14-0161

Dillow, E. (2013). *Narrative Discourse in Aphasia: Main Concept and Core Lexicon Analyses of the Cinderella Story*. Columbia: University of South Carolina.

Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H. (2018). Modeling confrontation naming and discourse informativeness using structural equation modeling. *Aphasiology*, 1–17. https://doi.org/10.1080/02687038.2018.1482404

Fergadiotis, G., Kellough, S., & Hula, W. D. (2015). Item response theory modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research*, *58*(3), 865–877.

Fromm, D. A., Forbes, M., Holland, A., & MacWhinney, B. (2013). PWAs and PBJs: Language for describing a simple procedure.

Goodglass, H., & Wingfield, A. (1997). Word-finding deficits in aphasia: Brain—behavior relations and clinical symptomatology. In *Anomia* (pp. 3–27). Elsevier.

Guo, Y. E., Togher, L., & Power, E. (2014). Speech pathology services for people with aphasia: What is the current practice in Singapore? *Disability and Rehabilitation*, *36*(8), 691–704. https://doi.org/10.3109/09638288.2013.804597

Herbert, R., Best, W., Hickin, J., Howard, D., & Osborne, F. (2003). Combining lexical and interactional approaches to therapy for word finding deficits in aphasia. *Aphasiology*, *17*(12), 1163–1186.

Herbert, R., Hickin, J., Howard, D., Osborne, F., & Best, W. (2008). Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? *Aphasiology*, *22*(2), 184–203.

Hickin, J., Herbert, R., Best, W., Howard, D., & Osborne, F. (2007). Efficacy of treatment: effects on word retrieval and conversation. In & C. P. S. Byung, K. Swinburn (Ed.), *Aphasia therapy file* (pp. 69–82).

Kambanaros, M. (2010). Action and object naming versus verb and noun retrieval in connected speech: Comparisons in late bilingual greek-english anomic speakers. *Aphasiology*, *24*(2), 210–230. https://doi.org/10.1080/02687030902958332

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test*. Philadelphia: PA: Pro-ed.

Kertesz, A. (1982). *Western aphasia battery test manual*. Psychological Corp.

Kertesz, A. (2006). *Western Aphasia Battery–Revised (WAB-R) Pro-Ed*. *Austin, TX*.

Law, S.-P., Kong, A. P.-H., Lai, L. W.-S., & Lai, C. (2015). Effects of context and word class on lexical retrieval in Chinese speakers with anomic aphasia. *Aphasiology*, *29*(1), 81–100.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database*. MIT Press.

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., & Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiology*, *24*(6–8), 856–868. https://doi.org/10.1080/02687030903452632

Manning, L., & Warrington, E. K. (1996). Two routes to naming: A case study. *Neuropsychologia*, *34*(8), 809–817. https://doi.org/10.1016/0028-3932(95)00166-2

Mayer, J., & Murray, L. (2003). Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. *Aphasiology*, *17*(5), 481–497.

Pashek, G. V., & Tompkins, C. A. (2002). Context and word class influences on lexical retrieval in aphasia. *Aphasiology*, *16*(3), 261–286. https://doi.org/10.1080/02687040143000573

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.

Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language*, *54*(2), 199–227.

Verna, A., Davidson, B., & Rose, T. (2009). Speech-language pathology services for people with aphasia: A survey of current practice in Australia. *International Journal of Speech-Language Pathology*, *11*(3), 191–205. https://doi.org/10.1080/17549500902726059

Williams, S. E., & Canter, G. J. (1982). The influence of situational context on naming performance in aphasic syndromes. *Brain and Language*, *17*(1), 92–106.

Wilshire, C. E., & McCarthy, R. A. (2002). Evidence for a context-sensitive word retrieval disorder in a case of nonfluent aphasia. *Cognitive Neuropsychology*, *19*(2), 165–186.

CHAPTER 3

STUDY I

Measuring word retrieval in narrative discourse: Core lexicon in aphasia

ABSTRACT

Background

Discourse analysis procedures are time-consuming and impractical in a clinical setting. Critical to clinicians are simple and informative discourse measures that require minimal time and labor to complete. Many studies, however, have overlooked difficulties that clinicians face. We recently developed core lexicon lists for nouns, verbs, adjectives, and adverbs for two narrative discourse tasks with healthy control groups. Core lexicon lists consist of important lexical items required to produce coherently meaningful discourse in response to discourse tasks. Measuring core lexicon is useful for quantifying word retrieval impairments at the discourse level in clinical populations.

Aim

The purpose of the current study was to apply an age-based core lexicon list for nouns, verbs, adjectives, and adverbs for the wordless picture books *Good Dog Carl* (*GDC*; Day, 1985) and *Picnic* (McCully, 1984) and determine how well the lists measured linguistic impairments in persons with Aphasia (PWA).

Method

Lemma forms were extracted from 470 control participants who were divided into seven age groups. Twenty-five core lexicons were identified for four word classes (nouns, verbs, adjectives, and adverbs) among the seven age groups. Then, the nouns, verbs, adjectives, and

adverbs for each PWA (N =11) were compared to the core lexicon for their respective age group. Percent agreement was computed by comparing the number of total items within each list to the number of items that PWA produced. A Spearman's correlation coefficient was computed between the WAB-R AQ and the percent agreement for each word type for PWA.

Results

The percentage of agreement for each word type among the age cohorts ranged between 56% and 96%. Of the four word types, core verbs significantly correlated with the WAB-AQs for both discourse tasks. A post-hoc analysis found significant differences between fluent and non-fluent aphasia for core verbs.

Conclusions

Core lexicon analysis appears to be a practical way to capture impairments in word retrieval at the discourse level. Core verbs may be a better indicator to understand holistic language performances for PWA. Use of the core lexicon checklist can serve as an option to reconcile ecological validity with clinical usability.

INTRODUCTION

Persons with aphasia (PWA) have been defined as having an acquired language impairment which presents with deficits in word retrieval (Goodglass & Winfield, 1997). Traditionally, speech-language pathologists have focused on PWA's ability to retrieve words because it indicates disruptions in lexico-semantic and/or phonological representations (Dell, Chang & Griffin, 1999). In this sense, commonly used measures of language difficulties in clinical settings capture word retrieval impairments at the single word level.

PWA's word retrieval impairment is thought to be different between word classes, such as nouns and verbs, depending on the aphasia language profile. In research involving comparisons between nouns and verbs, verb deficits have been reported in individuals with agrammatic aphasia on single-word naming tasks and discourse tasks, whereas other aphasia subtypes, such as fluent aphasia, represent relative deficits in retrieving nouns (Bates, Chen, Tzeng, Li, & Opie, 1991; Camarazza & Hills, 1991; Chen & Bates, 1998; Kim & Thompson, 2000; Luzzatti & Chierchia, 2000; Schwartz, Saffran, & Marin, 1980). However, some researchers have argued that there are no clear dissociations between retrieving nouns and verbs across different aphasia types (Berndt, Mitchum, Haendiges, & Sandson, 1997; Jonkers & Bastiaanse, 1998; Matzig, Drunks, Masterson, & Vigliocco, 2009; Williams & Canter, 1987; Zingerser & Berndt, 1990).

During discourse production, word retrieval problems in PWA have proven to be more dynamic because contextual effects may influence retrieval processes at the discourse level (Basso, Razzano, Faglioni, & Zanobio, 1990; Williams & Canter, 1982; Wilshire & McCarthy, 2002). Relatively few studies have investigated PWA's ability to retrieve words by word class beyond the word level (Berndt & Haendiges, 2000; Kambanaros, 2010; Mayer & Murray, 2003; Pashek & Tompkins, 2002; Zingeser & Berndt, 1988). Contrasting findings have been reported, where some studies have shown that persons with anomic aphasia performed better on retrieving nouns than verbs (Pashek & Tompkins, 2002; Zingeser & Berndt, 1988), and the others found an opposite pattern (Berndt & Haendiges, 2000). These conflicting results highlight the differences in lexical retrieval at the word level and discourse level, indicating that lexical retrieval at the word level may not inform or predict lexical retrieval at discourse level. Therefore, a goal of the current study is to develop a quantitative measure of word retrieval ability in discourse

production that is clinically practicable. In the following sections, we briefly summarize existing discourse measures, and challenges with these measures that led to the current study. Then, we review core lexicon measures developed in previous literatures.

Discourse Analysis

Discourse is any natural form of language comprising utterances or phrases (Wright & Capilouto, 2012) and may be "the most elaborative linguistic activity" (Ska, Duong, & Joanette, 2004, p. 302). Due to the complexity of discourse processing, quantifying discourse production in clinical settings is a challenging task (Armstrong, 2000; Prins & Bastiaanse 2004).

To date, researchers have suggested a great deal of outcome measures to examine the amount of information provided in discourse such as correct information unit (CIU; Nicholas & Brookshire, 1992) and main concept (Nicholas & Brookshire, 1995), which are rule-based scoring measures. In keeping with Nicholas and Brookshire's idea (1995), Wright and colleagues developed a main event analysis, which is operationally defined as essential elements within the discourse (Capilouto, Wright, & Wagovich, 2005) and is discourse-task specific. Recently, multi-level approaches that include micro- and macro-linguistic assessments have received experimental attention from researchers because they provide a breadth of information on discourse ability (Marini, Andreetta, Del Tin, & Carlomagno, 2011; Sherratt, 2007; Wright & Capilouto, 2012).

Although such analyses have been applied to empirically investigate discourse abilities in PWA, application and usability in clinical settings have not been readily investigated to our knowledge. Maddy, Howell, and Capilouto (2015) examined the extent to which clinicians have used discourse analysis to evaluate PWA in clinical settings. In semi-structured interviews with nine clinicians, they found that external influences such as time constraints and lack of training

obstruct application and use of discourse analysis. For example, discourse analysis requires collecting, transcribing, and analyzing language samples. A trained clinician generally requires more than four times the actual length of the discourse sample just to complete the transcription process alone (Armstrong, Brady, Mackenzie, & Norrie, 2007; Boles & Bombard, 1998; Elia, Liles, Duffy, Coelho, & Belanger, 1994; Boles & Bombard, 1998). This timeframe excludes the time required for training and analysis; thus, making many analyses impractical for use in clinical settings.

In recent discussions on the topic of discourse outcome measures, several groups of researchers agree that discourse analysis requires arduous processes (de Riesthal & Diehl, 2018; Dietz & Boyle, 2018a, 2018b; Kintz & Wright, 2018; Kurland & Stokes, 2018; Wallace, Worral, Rose, & Dorze, 2018; Whitworth, 2018). Commenting on roadblocks of discourse analysis, they have raised their voices in pursuing clinical feasibility to extend the use of discourse outcome measures by reducing time and effort. For many years, there has been an increasing emphasis on evaluating discourse without transcribing (McNeil, Doyle, Fossett, Park, & Goda, 2001; Olness, Gyger, & Thomas, 2012; but see de Riesthal & Diehl, 2018). Along with the advantage of lessening the burden on clinicians, non-transcription discourse analysis may also permit clinicians to achieve reliability (McNeil et al., 2001).

Core Lexicon in Aphasia

In acknowledgement of these clinical barriers for discourse analysis, recently researchers have developed a lexicon-based analysis that does not require an arduous transcription process (Dalton & Richardson, 2015; Dillow, 2013; Fromm, Forbes, Holland, & MacWhinney, 2013; MacWhinney, Fromm, Holland, Forbes, & Wright, 2010). Lexicon is not only a critical aspect of communication but the building block of discourse (Kintz, Fergadiotis, & Wright, 2016).

Without access to the intended word, the ability to deliver a message may be reduced. Moreover, core lexicon, which is one such analysis currently in development, refers to the pivotal lexical items required to produce a semantically meaningful and coherent narrative (MacWhinney et al., 2010). As such, it can be expected that core lexicon production reflects the ability to access the target word (MacWhinney et al., 2010), and further, informational discourse performance (Andreetta, Cantagallo & Marini, 2012).

MacWhinney and colleagues (2010) introduced a core lexicon analysis for the Cinderella story by analyzing the discourse samples from 25 healthy participants and 24 PWA. They collected the discourse samples from AphasiaBank (MacWhinney, Fromm, Forbes, & Holland, 2011), a collaborative project whose goal is to develop a database of language samples from PWA. Participants told the Cinderella story after looking through a 25-page wordless picture book. The researchers used the Computerized Language Analysis program (CLAN; MacWhinney, 2000) to extract the core lexicons from the language samples. They found that the PWA's discourse abilities were characterized with reduced lexical diversity and greater use of light verbs (i.e. frequently occurring verbs in language samples such as *be, have, come* etc…) compared to the control group. However, the core lexicon lists only included nouns and verbs, and the researchers did not consider other word classes, such as adjectives and adverbs, which may contribute to increased lexical diversity in discourse production (Sarno, Postman, Cho, & Norman, 2005).

Dalton and Richardson (2015) reported that a 24-item core lexicon list, independent of word class (i.e., verbs, nouns, adverbs, adjectives, etc.), discriminated between neurologically healthy controls and PWA. To develop the core lexicon list, the researchers accessed the transcripts of 92 healthy controls from Aphasia Bank. They extracted all the lemmas produced

within one of the discourse tasks, a sequential picture description task. The lemmas were extracted by using the CLAN command, where 24 lemmas produced by 50% or more of the control participants were included within the core lexicon. To determine if core lexicon could distinguish between the two groups, the researchers examined the transcripts of 166 healthy controls and 235 PWA. The researchers found a significantly different number of core lexicon items between PWA and healthy controls, and Broca's aphasia and other aphasia subtypes. They also concluded that the core lexicon list can reflect the participants' ability to convey the gist of a narration. However, the relative influence of lexical processing, known to be susceptible to aging, was not considered during development of the core lexicon lists.

Fromm, Forbes, Holland, and MacWhinney (2013) compared the core lexicon lists for a different type of discourse – procedural discourse (how to make a peanut butter & jelly sandwich). No differences were found between healthy controls (n = 145) and PWA (n = 141). They included additional measures as well: the number of words, the number of utterances, time on task, and mean length of utterance. The healthy control group produced significantly more words and utterances, and also had longer utterance durations compared to the aphasia group. Fromm and colleagues suggested these measures reflect quantitative differences among the groups. Further, they suggested that core lexicon is a qualitative assessment; in turn, suggesting the groups' procedural discourse samples differed quantitatively but not qualitatively. Results from these studies demonstrate potential pitfalls to using procedural discourse tasks for developing core lexicon measures such as fewer lexical items produced (Fergadiotis, Wright, & Capilouto, 2011).

Dillow (2013) analyzed the core lexicon lists for the Cinderella story. Dillow created core verb and noun lists following MacWhinney et al.'s (2010) procedures. In contrast to earlier

studies, they attempted to add an adjective core lexicon list, but they did not include it due to their criterion to establish the lexicon. The scores for core verbs, core nouns, and the entire core lexicon lists differentiated aphasia subtypes from the control group. They also analyzed how different word types of core lexicon affected the ability to differentiate the aphasia subtype groups. Core verbs differed for the following groupings (Anomic > Conduction > Wernicke > Broca): adults with anomic aphasia and adults with Broca's aphasia; adults with anomic aphasia and adults with conduction aphasia; and adults with Broca's aphasia and adults with Wernicke's aphasia. For core nouns, participants with anomic aphasia produced significantly more core nouns than those with Broca's and Wernicke's aphasia. Likewise, participants with conduction aphasia also produced significantly more core nouns than those with Broca's aphasia and Wernicke's aphasia. When considering the complete lexicon, adults with Broca's aphasia differed significantly from the adults with anomic aphasia and conduction aphasia. Compared to studies using an aggregated core lexicon list, this study demonstrates that separate core lexicon lists by word class differentiate each subtype from one another.

These findings are promising in that core lexicon analysis provides an alternative approach to more time-intensive, lexical-level discourse analyses. Whereas the transcription process and training are necessary for existing measures, clinicians can simply check if the words are present or not while listening to the recorded language samples once the core lexicon lists are established. However, limitations of previous research exist that need to be addressed. In generating the core lexicon, many researchers disregarded some word types (e.g. adjectives and adverbs) (Dillow, 2013; Fromm et al., 2013; MacWhinney et al., 2010) or combined words types to create a single core lexicon list (Dalton & Richardson, 2015). Different words types, such as nouns, verbs, adjectives, and adverbs, carry important and unique semantic information that

differentiate them (Neville, 2014). Based on previous research that production of modifiers manifested qualitative changes in language usage for persons with aphasia (Sarno et al., 2005), it would be worth developing core lexicon lists for different word types as an exploratory purpose. Moreover, previous studies of core lexicon have not considered age differences. Lexical selection by someone in their 20s may differ from someone in their 80s. Age should be considered when creating a core lexicon for a stimulus.

The purpose of the current study was to apply an age-based core lexicon list for nouns, verbs, adjectives, and adverbs for the wordless picture books, *Good Dog Carl* (*GDC*; Day, 1985) and *Picnic* (McCully, 1984), to determine how well the lists measure linguistic impairment in PWA. The specific aims of the current study, then, were two-fold: (1) determine the percent agreement between groups and their core lexicon and (2) examine the correlation among lexicon lists and aphasia impairment as determined by the aphasia quotient (AQ) from the Western Aphasia Battery – Revised (WAB-R; Kertesz, 2006). Based on the well-documented word retrieval deficits on verbs and nouns in PWA, we hypothesized that core nouns and verbs would positively correlate with the WAB-R AQs. If PWA demonstrate improved production of modifiers with better language performance as shown by Sarno and colleagues (2005), then it would be hypothesized that core adjectives and adverbs positively correlate with aphasia severity.

METHOD

Participants

Language samples from 470 cognitively healthy participants (273 females, 197 males) and 11 PWA were included in the study. The normative data presented are a subset of data from

a larger study examining discourse processing across the lifespan (Wright & Capilouto, 2017) and was approved by the respective universities (Arizona State University and University of Kentucky). The database included discourse samples and cognitive measures collected from over 470 participants ranging in age from 20 to 89 years. Control participants were divided into seven age groups (20s, 30s, 40s, 50s, 60s, 70s, and 80s). All control participants (a) were native English speakers, (b) passed hearing (Davis & Silverman, 1978) and vision screenings (Beukelman & Mirenda, 1998), (c) presented with normal cognitive functioning as indicated by the Mini-Mental State Exam (Folstein, Folstein, & McHugh, 2002), and (d) self-reported no history of stroke, head injury, or progressive neurogenic disorders. Demographic information for the control participants can be found in Table 3.1.

All PWA met the following 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 Hz, 1000 Hz, and 2000 Hz, (d) no reported history of other neurological disorders, (e) presented with aphasia as determined by performance on the WAB-R AQ subtests (Kertesz, 2006), (f) chronic aphasia (at least 6 months post-onset), and (g) left hemisphere damage. Initially, 13 PWA were recruited, and then 2 aphasia participants (P2 and P8) were disqualified from the study due to other neurological disorders. Thus, 11 right-handed participants with present or past evidence of stroke participated in this study. Demographic information for the PWA can be found in Table 3.2.

Experimental Procedures

All participants were tested individually in a laboratory setting. Since the normative data were collected for a large study, the cognitively healthy participants attended two sessions,

lasting no more than 2 hours for each session. Prior to study participation, they completed consent forms, and then completed screening measures to confirm that they met the inclusion criteria. Next, a cognitive test battery and a set of discourse tasks were administered. Order of test administration was randomized across participants. The cognitive test and discourse task results irrelevant to this study are not reported here.

For participants in the PWA group, the WAB-R was administered first and then cognitive and discourse tasks were randomized across participants. During the experimental procedures, they were allowed to take breaks as needed. This study is focused solely on some of those discourse measures (described below).

Discourse Task

Two wordless picture books were used to collect narrative discourse samples from participants. They included *Good Dog Carl* (*GDC*; Day, 1985) and *Picnic* (McCully, 1984). Because limited to no text is included in the books, the task is a storytelling or story generation task, rather than a story-retelling task. Storytelling tasks are "more representative of spontaneous communication" (Liles, 1993, cited in Hughes, McGillivray, & Schmidek, 1997, p. 19). Additionally, because participants are telling stories from books rather than from shorter pictured stimuli (e.g., single pictures), participants provide longer samples and use a more diverse vocabulary (Fergadiotis & Wright, 2011; Fergadiotis et al., 2011; Wright, Capilouto, Srinivasan, & Fergadiotis, 2011). *GDC* is a 30-page book that follows a temporally-driven story structure conveying the events that unfold as a dog is left to take care of a baby. The story *Picnic* is a 31-page story that represents a spatially- and temporally-driven story structure conveying the adventures of a family of mice going on a picnic. For the discourse task, the examiner provided the following instructions, "These are wordless picture books that allow an individual to make-

up their own story. First, I'll look through the book to get an idea of the story." Then, the examiner provided an example to participants with another story, for *The Great Ape* (Krahn, 1978). Finally, 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. While telling the story, the books were still viewable by the participant.

Language Sample Preparation

All samples were either audio or video recorded, and then orthographically transcribed by trained research assistants using a set of programs called CLAN.

Ten percent of the control participants were randomly selected for inter-rater and intra-rater reliability for the entire transcription. The inter- and intra-rater agreements were 95% and 98%, respectively. For the aphasia group, two persons with aphasia were selected due to the small number of participants, and inter- and intra-rater agreements for the entire transcription were 91% and 93%, respectively.

Core Lexicon

The core lexicons for *GDC* and *Picnic* were created by determining the 25 most frequently used lemmas produced for nouns, verbs, adjectives, and adverbs within each age group. The researchers accomplished this goal by assigning the proper syntactic category for each word within the narrative transcripts and extracting the lemma forms and their respective frequencies within each age group. To assign proper syntactic categories, the researchers used CLAN program (Child Language Analysis; MacWhinney, 2000) and the methods outlined by MacWhinney and colleagues (MacWhinney, 2000; MacWhinney et al., 2010). CLAN uses the programs MOR and POST, which are respectively tied to a dictionary of lexical items and English grammar rules, to automatically categorize words into their respective syntactic category

with an accuracy of 95% (for review see, MacWhinney, 2000). Once the words within each transcript were categorized, the researchers automatically extracted *GDC* and *Picnic* into separate files using the GEM program of CLAN. MOR, POST, and GEM are terms for CLAN commands. The MOR command is used primarily for morphosyntactic analysis for each word. The POST command following the MOR command automatically resolves grammatical ambiguity. The GEM command is to sort different discourse tasks in the transcripts (see Appendix 3.A for the CLAN commands). This step is necessary to create two independent lists for each story. For each story and age group, the lemma forms were extracted for all the participants into a single list that included their frequency information. For example, if 20 participants used the lemma *go* once and a single participant used *go* 5 times, the lemma list would indicate that *go* was produced 25 times for that age group. The top 25 most frequent lemmas were collected for each word class within each story for every age group. While the top 25 most frequent lemmas are an arbitrary cut-off, previous researchers used similar numbers (Dalton & Richardson, 2015). A complete list of the top 25 core lexicon for each age group for the two stimuli is presented in Appendix 3.B.

Percent agreement was determined by comparing the 25 core lexical items within each list among seven neurologically healthy groups. Percent agreement was calculated by dividing the number of agreements by the total number of core lexical items on each list (the number of agreements / 25) * 100. For example, an aphasia speaker (P1) who is in his 60's was evaluated by using the 60's age group core lexicon lists. If P1 produces 4 items from the 60's core verb list, the numerator is 4 and the denominator is 25 in the fraction.

Core Lexicon Production in Aphasia

The PWA's transcripts were prepared for analysis in a similar manner as described earlier. Counting of how many core lexical items were produced in PWA was based on PWA's transcripts. These lists were compared to the age-matched core lexicon list for each story. For this study, we chose not to count synonyms, to maintain consistency with Dalton and Richardson's (2015) procedures, which acknowledges the importance of producing the target words (e.g., Andreetta et al., 2012; Verhaegen & Poncelet, 2013). If a PWA produced any lemmas on any of the core lexicon lists, they would receive a point. If the PWA did not produce the lemma form, they did not receive a point. Only one point was provided regardless of how many times the lemma form may have been used by the participant. The number of lemmas produced was divided by the total number of lemmas on the core lexicon list for each syntactic category type resulting in a percent agreement between the PWA and age-matched cohorts for the core lexicon lists.

RESULTS

The purpose of the study was to apply a core lexicon list for nouns, verbs, adjectives, and adverbs within the narrative discourse, *GDC* and *Picnic*, for different age groups and compare to core lexicon productions by PWA. These age-based core lexicon lists were utilized to address the aims of the current study.

The percentage of agreement for each word type was calculated across the seven age groups. Adverbs appeared to have the best agreement among the age groups with the lowest agreement only at 72%. Verbs had the next best agreement among the age cohorts, ranging between 64% to 92%. The percent agreement for adjectives ranged from 56% to 92%. The percent agreement for nouns ranged from 56% to 98%. See Tables 3.3 – 3.6 for agreements

among age groups for syntactic category types for each narrative discourse task (*GDC* and *Picnic*).

To investigate the relationship between the core lexicon and aphasia impairment, the percent agreement for each word class was obtained between the PWA and age-matched cohorts for the core lexicon lists. Spearman's correlation coefficients were computed between WAB-R AQs and core lexicon agreements for nouns, verbs, adjectives, and adverbs for each narrative task. For both *GDC* and *Picnic*, significant correlations were found between core verbs and WAB-AQs, $r(9) = .869$, $p < .001$, $r(9) = .892$ $p < .001$. PWA with better AQs had greater core lexicon agreements for verbs. Significant correlations were not found among AQs and other word classes (nouns, adjectives and adverbs) (See Tables 3.7 & 3.8).

Post-hoc analysis: aphasia type

A post hoc analysis was conducted to determine if production of different word types obtained by the core lexicon measure differed between individuals with fluent (N = 5) and individuals with non-fluent (N = 6) types of aphasia. To conduct this analysis, the PWA were divided into two groups (fluent vs. non-fluent) based on the WAB-R aphasia classification. A Mann-Whitney *U* test indicated that for *GDC*, production of core verbs was significantly greater for fluent aphasia (Mean Rank = 7.50) than for non-fluent aphasia (Mean Rank = 3.50), $U = 2.50$, $z = -2.11$, $p < .05$. For *Picnic*, fluent aphasia (Mean Rank = 8.30) also produced more core verbs than non-fluent aphasia (Mean Rank = 4.08), $U = 000$, $z = -2.124$, $p < .05$ (See Table 3.9).

DISCUSSION

The purpose of the study was to apply age-developed core lexicon lists for the narrative discourse tasks *GDC* and *Picnic* to determine if core lexicon lists for word type would correlate

with aphasia severity. For the normative data, while comparatively high agreement across age groups was observed for adjectives and verbs, adverb and noun use had considerable variability across the age cohorts, suggesting a need to develop and use core lexicon lists that account for age with clinical populations. Further, only verbs significantly correlated with WAB-AQs for both narrative tasks for the PWA. These findings suggest that the core lexicon comparisons between age-matched controls and PWA may be useful for determining atypical patterns of lexical usage in discourse production, which in turn is reflective of aphasia severity.

Core Lexicon and Aphasia

Core verbs for both tasks significantly correlated with overall aphasia severity as measured by the WAB-R AQ, providing partial support for our hypothesis that core verbs and nouns correlate with AQs. These findings agree with findings by other researchers who were able to differentiate aphasia subtypes (Dillow, 2011). Whereas some researchers have created a single core lexicon list (Dalton & Richardson, 2015; MacWhinney et al., 2010), Dillow (2011) demonstrated that a single list is not sufficiently able to discern between aphasia types and thus created core lexicon lists for nouns and verbs separately. Our study extended these results by adding lists for adverbs and adjectives as an exploratory investigation. In the current study, we did not have a large enough sample to consider different subtypes of aphasia and determine if each core lexicon list differed across aphasia subtypes. However, findings of the post-hoc analysis lend weight to our results in that the only difference identified was that individuals with fluent aphasia produced significantly more core verbs than individuals with non-fluent aphasia for both tasks.

Our results support and extend Dillow's (2011) results, wherein verbs are important in differentiating aphasia subtypes. As overall aphasia severity increases, fewer verbs are produced.

This finding is unsurprising, since verbs are often considered the building blocks or central themes of utterances (Healy & Miller, 1970). Additionally, these findings have critical implications in terms of how researchers and clinicians should assess and treat verbs in discourse production of PWA. However, it was somewhat surprising that no significant correlations were found between core noun production and overall language severity obtained from the standardized, norm-referenced measure (i.e., WAB-R AQ), considering the substantial impact of noun production in clinical decisions. For language assessment in PWA, the WAB-R and the Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 1976) require naming noun objects and are the most frequently used tests in clinical settings (Guo, Togher & Power, 2014; Verna, Davidson & Rose, 2009). Based on these findings, it may be insufficient for clinicians to rely on performance of noun production alone when drawing clinical decisions regarding word retrieval abilities of their patients with aphasia.

Further, the initial analyses based on 11 PWAs did not demonstrate significant correlations between aphasia severity and production of adjectives and adverbs. However, the subsequent statistical analyses (*Mann-Whitney U-Test*) of the fluent and non-fluent aphasia groups detected lower statistical power (p = .052) for capturing significant differences in adverb production (See Table 9). Because of the small number of aphasia participants and relatively restricted range of aphasia severity included in the separate group analyses, these results should be interpreted with caution (i.e., fluent aphasia groups presented with more mild aphasia compared to non-fluent aphasia). Future studies should consider potential joint effects of different word types to capture the level of aphasia severity.

In contrast to the current study, previous investigators have employed different elicitation techniques such as story retelling of Cinderella (Dillow, 2011; MacWhinney, 2010), and

procedural discourse (Dalton & Richardson, 2015; Fromm et al., 2013). Considering that the core lexicon measure is developed based on the entire spoken lexicon, a sufficient number of words should be produced to create a reliable and sensitive measure for capturing unusual lexicon patterns of PWA. Although narrative discourse obtained from wordless picture books has not been used in clinical settings frequently, it does provide lexically diverse language samples (Fergadiotis & Wright 2011), thereby increasing the probability of capturing the severity of aphasia using this measure. Additionally, the existence of pictorial stimuli may be an important factor to elicit discourse samples with high quality and quantity (Grosjean, 1980). A task that provides picture stimuli having frame-by-frame presentation may function as cognitive schema, which leads to more episodes (Coelho, 2002). As such, narrative discourse tasks with pictorial support may be appropriate for collecting language samples, as well as for developing core lexicon measures.

There is no converging evidence from previous research with respect to criterion for the lemmas. For example, MacWhinney and colleagues (2012) did not stipulate a criterion and generated 10 core nouns and 10 core verbs. Fromm and colleagues (2013) did not specify an inclusionary criterion and included 10 core lexicon items as well, though with comparatively short language samples obtained from procedural discourse. Other studies required that at least 50% of the core lexical items be produced by the control participants to be included in the core lexicon list (Dalton, & Richardson, 2015; Dillow, 2011). Given that the core lexicon measure is a relatively novel method, an important next step is to determine the impact of different inclusionary criteria for lexical items, and then investigate the sensitivity and specificity of measuring language impairments. For best practice and usability of the core lexicon measure, a systematic approach to the criterion should be considered in future investigations.

Clinical Implications

Discourse outcome measures are evolving in response to clinical utility. Such changes can enhance our understanding of discourse impairment of PWA, as well as aid in alleviating some difficulties that clinicians face. This study is a step forward in addressing the issue of clinical feasibility for discourse analysis in clinical settings. Researchers investigating discourse ability in PWA claim that the transcription process is an obstacle that prevents clinicians from using discourse analysis in clinical settings (de Riesthal & Diehl 2018; Kintz & Wright 2018;).

In this sense, the core lexicon measure is a meaningful outcome because it is easily quantifiable and time-saving for assessment without transcription. Additionally, eliciting sufficient quality and quantity of language samples in a limited period of time is important. The explicit task instructions (identified in the method) are distinct from traditional instructions (i.e., "tell me everything you see going on in this picture") and induce individuals to provide the core event line of pictures depicted in narrative discourse (Olness 2006; Wright & Capilouto, 2009). In this study, we did not ask participants to describe every scene, but instead to build the story. This led participants to focus on temporal and/or causal information, not simply list all objects viewed in each scene. Most participants in this study took between 5 to 15 minutes to complete both tasks, whereas one PWA with the longest language sample for our participants took 21 minutes to finish them. In turn, we were able to elicit a language sample in a very appropriate time frame despite comparatively more picture stimuli included in the task.

Another clinical contribution of the current study is that the core lexicon measure was created based on the performance of cognitively healthy controls. Discourse disruptions featured in PWA lie on the continuum of normal discourse performance. By contrasting PWA's lexical usage to typical lexical usage produced by cognitively healthy controls within similar age

cohorts, we can gain some insight into the nature of PWA's language profiles and to what extent they are preserved or impaired. Finally, though separate core lexicon lists by word class may be useful for evaluating overall changes in lexical use before and after treatment, they are restricted to providing clinical information about lexical retrieval. Core lexicon does not inform, clinically, about syntactic structure, rate of speech, or fluency.

CONCLUSIONS AND FUTURE DIRECTIONS

Multiple core lexicon lists were developed in this study for two discourse elicitation tasks and seven, 10-year age-cohort groups and compared to narratives elicited from PWA to determine the suitability of core lexicons for predicting aphasia severity and potential clinical use. Results of the study are promising, as they broaden our understanding of how meaningful the verb core lexicon is for PWA and also have clinical implications. The core verbs were verified as a comparatively simple means for predicting the language function of PWA, while other core lexicon lists were not. These findings have potential clinical implications in that verb counts (i.e., using a core verb list or counting verbs produced) might be a discourse measure that is sensitive to capturing comprehensive language ability.

However, there are several limitations to the study that need to be considered in future investigations. A major limitation is the construction of the core lexicon lists. The lists included the 25 most common lemmas produced by cognitively healthy participants for each stimulus. While there is a precedent for defining text from a frequency list (Gottron, 2009), the cut-off was mostly arbitrary with ease of use being the most important factor in that decision. A combined frequency list may be plagued with outliers if an individual uses a single lemma significantly more than others. For example, a discourse sample that includes the word "no" a thousand times

might place “no” at the top of the frequency list, but it would not be descriptive of the text. Future research should address this issue.

We were not able to find noun, adjective, or adverb core lexicon lists that were sensitive to severity of aphasia, perhaps due to the small sample size. Reviewing the demographic information of these participants with aphasia, nearly half presented with Broca's aphasia. Possibly, the expected, limited verb retrieval of individuals with Broca's aphasia drove the statistically significant results. In the same vein, more fluent types of aphasia need to be included to ensure the necessity of expanding grammatical category for both research and clinical judgment similar to how Sarno and colleagues (2005) were able to demonstrate the predictive value of the production of modifiers. Future studies should include a larger number of participants as well as a sufficient number of participants with different aphasia types so as to determine whether the findings are specific to type of aphasia.

Lastly, core lexicon lists should be applied to different discourse elicitation tasks such as picture descriptions, procedural discourse tasks, and storytelling, to explore discourse adequacy by measuring the core lexicon that is most useful for clinical populations. Finally, to establish ecological validity and utility of the core lexicon measure, it is essential that researchers investigate its correlations to other linguistic measures as well as to the standardized tools. Along with the acceptable validity, it could be expected that such an effort will acquire a more useful clinical prediction by requiring less time, training, and efforts when completing these key evaluations in clinical settings.

Acknowledgements

This research was partially supported by National Institute on Aging Grant R01AG029476. We are especially grateful to the study participants.

References

Andreetta, S., Cantagallo, A., & Marini, A. (2012). Narrative discourse in anomic aphasia. *Neuropsychologia*, *50*, 1787–1793. https://doi.org/10.1016/j.neuropsychologia.2012.04.003

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

Armstrong, L., Brady, M., Mackenzie, C., & Norrie, J. (2007). Transcription- less analysis of aphasic discourse: A clinician's dream or a possibility? *Aphasiology*, *21*(3–4), 355–374.

Basso, A., Razzano, C., Faglioni, P., & Zanobio, M. E. (1990). Confrontation naming, picture description and action naming in aphasic patients. *Aphasiology, 4*(2), 185–195.

Bates, E., Chen, S., Tzeng, O., Li, P., & Opie, M. (1991). The noun-verb problem in Chinese aphasia. *Brain and language, 41*(2), 203-233.

Berndt, R. S., & Haendiges, A. N. (2000). Grammatical class in word and sentence production: Evidence from an aphasic patient. *Journal of Memory and Language*, *43*(2), 249–273.

Berndt, R.S., Mitchum, C.C., Haendiges, A.N. & Sandson, J. (1997). Verb retrieval in aphasia. 1. Characterizing single word impairments. *Brain and language, 56*(1), 68-106.

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

Boles, L., & Bombard, T. (1998). Conversational discourse analysis: Appropriate and useful sample sizes. *Aphasiology, 12*(7–8), 547–560.

Capilouto, G., Wright, H. H., & Wagovich, S. A. (2005). CIU and main event analyses of the structured discourse of older and younger adults. *Journal of Communication Disorders*, 38, 431–444. https://doi.org/10.1016/j.jcomdis.2005.03.005

Chen, S. & Bates, E. (1998). The dissociation between nouns and verbs in Broca's and Wernicke's aphasia: Findings from Chinese. *Aphasiology, 12*(1), 5-36.

Coelho, C. A. (2002). Story narratives of adults with closed head injury and non-brain-injured adults: influence of socioeconomic status, elicitation task, and executive functioning. *Journal of Speech, Language, and Hearing Research, 45*, 1232–1248. https://doi.org/10.1044/1092-4388(2002/099)

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.

Davis, H., & Silverman, S. R. (1970). *Hearing and deafness*. New York, NY: Holt, Rinehart & Winston of Canada Ltd).

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

de Riesthal, M., & Diehl, S. K. (2018). Conceptual, methodological, and clinical considerations for a core outcome set for discourse. *Aphasiology*, *32*(4), 469–471.

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, *23*(4), 517–542.

Dietz, A., & Boyle, M. (2018a). Discourse measurement in aphasia: consensus and caveats. *Aphasiology*, *32*(4), 487–492.

Dietz, A., & Boyle, M. (2018b). Discourse measurement in aphasia research: have we reached the tipping point? *Aphasiology*, *32*(4), 459–464.

Dillow, E. (2013). *Narrative Discourse in Aphasia: Main Concept and Core Lexicon Analyses of the Cinderella Story*. Columbia: University of South Carolina.

Elia, D., Liles, B. Z., Duffy, R. J., Coelho, C. A., & Belanger, S. A. (1994). An investigation of sample size in conversational analysis. In *ASHA Convention, New Orleans, LA*.

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology*, *25*(11), 1414–1430. https://doi.org/10.1080/02687038.2011.603898

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

Folstein, M., Folstein, S., & Fanjiang, G. (2001). *Mini-mental State Examination* – 2nd Edition. Lutz, FL: PAR.

Fromm, D. A., Forbes, M., Holland, A., & MacWhinney, B. (2013). PWAs and PBJs: Language for describing a simple procedure.

Gottron, T. (2009). Document word clouds: Visualising web documents as tag clouds to aid users in relevance decisions. In *International Conference on Theory and Practice of Digital Libraries* (pp. 94–105). Springer.

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

Guo, Y. E., Togher, L., & Power, E. (2014). Speech pathology services for people with aphasia: What is the current practice in Singapore? *Disability and Rehabilitation*, *36*(8), 691–704. https://doi.org/10.3109/09638288.2013.804597

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], 2017-03-02. http://doi.org/10.3886/ICPSR36634.v1

Healy, A.F., & Miller, G.A. (1970). The verb as the main determinant of sentence meaning. *Psychonomic Science, 20*(6), 372-372.

Hughes, D.L., McGillivray, L., & Schmidek, M. (1997). *Guide to narrative language: Procedures for assessment.* Eau Claire, WI: Thinking.

Jonkers, R., & Bastiaanse, R. (1998). How selective are selective word class deficits? Two case studies of action and object naming. *Aphasiology, 12*(3), 245-256.

Kambanaros, M. (2010). Action and object naming versus verb and noun retrieval in connected speech: Comparisons in late bilingual greek-english anomic speakers. *Aphasiology*, *24*(2), 210–230. https://doi.org/10.1080/02687030902958332

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test*. Philadelphia: PA: Pro-ed.

Kemmerer, D. (2005). The spatial and temporal meanings of English prepositions can be independently impaired. *Neuropsychologia*, *43*(5), 797–806.

Kertesz, A. (1982). *Western aphasia battery test manual*. Psychological Corp.

Kertesz, A. (2006). *Western Aphasia Battery–Revised (WAB-R) Pro-Ed*. *Austin, TX*.

Kim, M., & Thompson, C. K. (2004). Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization. *Brain and Language*, *88*(1), 1–20.

Kintz, S., Fergadiotis, G., Wright, H. H., & Wright, H. H. (2016). Aging effects on discourse production. *Cognition, Language, and Aging*, 81–106.

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

Kurland, J., & Stokes, P. (2018). Let's talk real talk: an argument to include conversation in a D-COS for aphasia research with an acknowledgment of the challenges ahead. *Aphasiology*, *32*(4), 475–478.

Liles, B.Z. (1993). Narrative discourse in children with language disorders and children with normal language: A critical review of the literature. *Journal of Speech, Language, and Hearing Research, 36*(5), 868-882.

Luzzatti, C., & Chierchia, G. (2002). On the nature of selective deficits involving nouns and verbs. *Italian Journal of Linguistics, 14*, 43-72.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database*. MIT 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. https://doi.org/10.1080/02687030903452632

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., Boewe, A., Caltagirone, C., & Carlomagno, S. (2005). Age-related differences in the production of textual descriptions. *Journal of Psycholinguistic Research*, *34*(5), 439–463. https://doi.org/10.1007/s10936-005-6203-z

Mätzig, S., Druks, J., Masterson, J., & Vigliocco, G. (2009). Noun and verb differences in picture naming: Past studies and new evidence. *Cortex*, *45*(6), 738–758.

Mayer, J., & Murray, L. (2003). Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. *Aphasiology*, *17*(5), 481–497.

McCully, E. A. (1984). *Picnic*. Harper & Row New York.

McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology*, *15*(10–11), 991–1006. https://doi.org/10.1080/02687040143000348

Neville, H.J. (2014). Procedings from a Conference to Honor Alvin M. Liberman: *Language Hemisphere? Modularity and the Motor Theory of Speech Perception.* Psychology Press.

Olness, G.S. (2006). Genre, verb, and coherence in picture-elicited discourse of adults with aphasia. *Aphasiology*, *20*(02-04), 175-187.

Olness, G. S., Gyger, J., & Thomas, K. (2012). Analysis of Narrative Functionality: Toward Evidence-based Approaches in Managed Care Settings. *Seminars in Speech and Language*, *33*, 55–67. https://doi.org/10.1055/s-0031-1301163

Pashek, G. V., & Tompkins, C. A. (2002). Context and word class influences on lexical retrieval in aphasia. *Aphasiology*, *16*(3), 261–286. https://doi.org/10.1080/02687040143000573

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

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.

Schwartz, M.F., Saffran, E.M., & Marin, O.S. (1980). The word order problem in agrammatism: I. Comprehension. *Brain and language, 10*(2), pp. 249-262.

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

Ska, B., Duong, A., & Joanette, Y. (2004). Discourse impairments. In R. D. Kent (Ed.), *The MIT encyclopedia of communication disorder* (pp. 302-304). Cambridge, MA: The MIT Press.

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.

Verna, A., Davidson, B., & Rose, T. (2009). Speech-language pathology services for people with aphasia: A survey of current practice in Australia. *International Journal of Speech-Language Pathology*, *11*(3), 191–205. https://doi.org/10.1080/17549500902726059

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 standardisation of discourse measures? *Aphasiology*, *32*(4), 479–482.

Whitworth, A. (2018). The tipping point: are we nearly there yet? *Aphasiology*, *32*(4), 483–486.

Williams, S. E., & Canter, G. J. (1982). The influence of situational context on naming performance in aphasic syndromes. *Brain and Language*, *17*(1), 92–106.

Wilshire, C. E., & McCarthy, R. A. (2002). Evidence for a context-sensitive word retrieval disorder in a case of nonfluent aphasia. *Cognitive Neuropsychology*, *19*(2), 165–186.

Wright, H. H., & Capilouto, G. J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology*, *23*(10), 1295–1308. https://doi.org/10.1080/02687030902826844

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

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)

Zingeser, L. B., & Berndt, R. S. (1988). Grammatical class and context effects in a case of pre-anomia: implications for models of language processing. *Cognitive Neuropsychology*, *5*, 473–516.

Table 3.1

Neurologically Healthy Adult Demographic Information

Age Group	N (F:M)	Age (SD)	Education (SD)
20s	66 (35:31)	23.93 (3.69)	15.76(1.93)
30s	63 (39:24)	34.12 (3.11)	16.15(3.28)
40s	67 (41:26)	44.34 (3.01)	15.36 (2.54)
50s	68 (43:25)	55.57 (2.65)	15.85 (2.54)
60s	67 (38:29)	64.78 (4.93)	15.45 (2.49)
70s	76 (43:33)	73.85(2.88)	15.32(2.32)
80s	63 (34:29)	83.29 (2.71)	14.76 (2.70)
Total	470 (273:197)		

Table 3.2

Participants with Aphasia Demographic Information

	Age	Gender	Education	WAB-R AQ	Post-onset (month)	Aphasia Type (Fluent/Non-fluent)
P1	65	M	18	76.3	67	Conduction (Fluent)
P3	73	M	12	85.2	25	Anomic (Fluent)
P4	84	F	12	62.6	26	Conduction (Fluent)
P5	55	M	14	57.6	26	Broca's (Non-fluent)
P6	66	F	14	56.3	171	Broca's (Non-fluent)
P7	34	F	14	90.7	21	Anomic (Fluent)
P9	38	F	14	57.7	151	Broca's (Non-fluent)
P10	62	F	20	61.3	96	Broca's (Non-fluent)
P11	72	M	12	64.9	57	Transcortical motor (Non-fluent)
P12	65	F	11	89.4	120	Anomic (Fluent)
P13	65	M	14	54.4	N/A	Broca's (Non-fluent)
Average (SD)	61.7 (14.7)		14.1 (2.7)	68.8 (14)		

Note. WAB-R AQ = Western Aphasia Battery-Revised (Kertesz, 2006). Maximum WAB-R AQ raw score = 100.

Table 3.3

Percent agreement between cognitively-healthy age cohorts for Nouns

	20s	30s	40s	50s	60s	70s	80s
20s		80%	72%	80%	60%	60%	60%
30s	88%		80%	68%	80%	72%	68%
40s	92%	92%		76%	72%	64%	60%
50s	84%	92%	92%		80%	60%	56%
60s	80%	84%	88%	98%		68%	64%
70s	72%	80%	80%	84%	88%		72%
80s	68%	72%	72%	80%	80%	84%	

Note. GDC appears in the section of bottom left. Picnic with grey coloring appears in the section of upper right.

Table 3.4

Percent agreement between cognitively-healthy age cohorts for Verbs

	20s	30s	40s	50s	60s	70s	80s
20s		80%	88%	88%	80%	76%	72%
30s	92%		72%	76%	68%	68%	64%
40s	80%	84%		84%	92%	84%	76%
50s	88%	84%	80%		80%	80%	68%
60s	88%	80%	80%	84%		88%	84%
70s	88%	80%	76%	80%	88%		80%
80s	76%	72%	80%	76%	80%	76%	

Note. GDC appears in the section of bottom left. Picnic with grey coloring appears in the section of upper right.

Table 3.5

Percent agreement between cognitively-healthy age cohorts for Adjectives

	20s	30s	40s	50s	60s	70s	80s
20s		84%	80%	76%	72%	80%	76%
30s	72%		84%	80%	76%	80%	72%
40s	56%	72%		88%	88%	92%	84%
50s	64%	76%	68%		88%	88%	84%
60s	60%	80%	68%	72%		92%	84%
70s	60%	76%	64%	80%	76%		92%
80s	64%	76%	64%	80%	76%	76%	

Note. GDC in the section of bottom left. Picnic with grey coloring appears in the section of upper right.

Table 3.6

Percent agreement between cognitively-healthy age cohorts for Adverbs

	20s	30s	40s	50s	60s	70s	80s
20s		84%	80%	76%	72%	80%	76%
30s	80%		84%	80%	76%	80%	72%
40s	76%	84%		88%	88%	92%	84%
50s	80%	88%	84%		88%	88%	84%
60s	80%	88%	84%	84%		92%	84%
70s	80%	84%	88%	96%	84%		92%
80s	80%	88%	88%	88%	88%	92%	

Note. GDC appears in the section of bottom left. Picnic with grey coloring appears in the section of upper right.

Table 3.7

Percent Agreement for the participants with aphasia with their respective age group for the Core Lexicon

Participant		Good Dog Carl				Picnic			
ID	Age group	Nouns	Verbs	Adjectives	Adverbs	Nouns	Verbs	Adjectives	Adverbs
P1	60s	36	16	8	4	24	28	28	28
P3	70s	24	44	16	16	16	28	28	24
P4	80s	44	32	36	40	24	16	44	36
P5	50s	36	8	20	16	16	20	8	16
P6	60s	N/A	N/A	N/A	N/A	16	12	8	8
P7	30s	56	48	16	20	36	52	24	16
P9	30s	48	20	16	24	16	16	16	20
P10	60s	40	16	16	12	24	8	12	4
P11	70s	8	28	28	32	8	28	28	32
P12	60s	52	48	20	20	52	48	28	40
P13	60s	48	8	24	24	28	4	32	12

Note. Good Dog Carl (Day, 1985); Picnic (McCully, 1984).

Table 3.8

Correlations (Spearman's rho) between AQs and core lexicon by word class

Comparison	Spearman's Rho	Significance
GDC		
Nouns	.146	.687
Verbs	.869**	.001
Adjectives	-.307	.388
Adverbs	-.171	.636
Picnic		
Nouns	.338	.309
Verbs	892**	.000
Adjectives	.266	.429
Adverbs	.574	.065

Note. Good Dog Carl (Day, 1985); Picnic (McCully, 1984)
*p<.05, **p<.01

Table 3.9

Mann-Whitney U Test of difference in the core lexicon between two aphasia types (fluent vs. non-fluent)

Task		n	Mean Rank	*p*
GDC	Nouns			
	Fluent	5	6.50	.310
	Non-fluent	5	4.50	
	Verbs			
	Fluent	5	7.50	.032*
	Non-fluent	5	3.50	
	Adjectives			
	Fluent	5	5.20	.841
	Non-fluent	5	5.80	
	Adverbs			
	Fluent	5	5.10	.690
	Non-fluent	5	5.90	
Picnic	Nouns			
	Fluent	5	8.00	.082
	Non-fluent	6	4.33	
	Verbs			
	Fluent	5	8.30	.030*
	Non-fluent	6	4.08	
	Adjectives			
	Fluent	5	7.70	.126
	Non-fluent	6	4.58	
	Adverbs			
	Fluent	5	8.10	.052
	Non-fluent	6	4.25	

Note. Good Dog Carl (Day, 1985); Picnic (McCully, 1984)
*$p < .05$, **$p < .001$

Appendix 3.A: The CLAN commands

(1) Generate a morphological Analysis*: mor +t*SUB *.gem.cex*
(2) Generate Syntactic Categories: *post +t*SUB *.gem.mor.cex*
(3) Extract the two stories: *gem +t%mor +t*SUB +sStory +d +f _n *.cha*
(4) Extract all lemma forms with frequencies information: *freq +t%mor +s@"r*,|*,o%" +u +d2 *.gem.mor.pst.cex*

Appendix 3.B: The top 25 core lexicon produced by the control group.

Good Dog Carl

20s				**30s**			
Nouns	**Verbs**	**Adjectives**	**Adverbs**	**Nouns**	**Verbs**	**Adjectives**	**Adverbs**
baby	go	good	then	baby	go	little	then
dog	get	little	back	Carl	get	good	back
Carl	put	back	there	dog	put	clean	all
mother	look	hungry	just	crib	take	back	there
mom	take	big	all	mom	make	hungry	up
crib	come	messy	where	bed	run	big	upstairs
bed	have	open	upstairs	back	have	open	now
back	play	sure	up	laundry	come	great	away
room	make	own	out	room	decide	sure	where
laundry	run	left	down	mother	play	nice	out
mess	see	dirty	away	milk	see	happy	just
chute	decide	happy	next	time	watch	ready	so
milk	leave	nice	very	mess	clean	messy	next
bread	say	huge	in	window	leave	dirty	on
home	clean	next	now	bread	dance	tired	down
window	eat	right	again	cookie	jump	left	off
time	let	able	also	chute	turn	next	very
child	turn	hot	around	home	ride	smart	in
butter	dry	dangerous	so	butter	let	first	shortly
cookie	do	first	shortly	kitchen	dry	right	sure
bath	know	great	on	house	say	whole	again
kitchen	watch	long	more	chocolate	know	awesome	over
grape	find	ready	how	grape	eat	huge	really

fishtank	ride	young	once	bath	wash	old	around
soap	dance	early	as	floor	find	pretty	soon

40s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
baby	go	little	then
Carl	get	good	back
dog	put	hungry	there
crib	look	back	up
back	have	big	all
bed	take	clean	just
mom	come	dry	now
laundry	play	messy	upstairs
mother	run	great	out
milk	make	sure	very
room	see	dirty	where
chute	let	happy	away
window	know	nice	down
mess	say	fun	in
bread	clean	old	on
chocolate	watch	next	around
time	ride	ready	here
cookie	decide	right	over
floor	leave	first	off
kitchen	do	safe	next
soap	open	whole	again
butter	climb	wonderful	right
home	find	bad	apparently

50s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
baby	go	little	then
Carl	get	good	back
dog	look	big	there
bed	put	hungry	up
mother	take	sure	all
crib	have	messy	just
back	clean	open	now
mom	come	next	upstairs
laundry	play	fun	out
milk	see	dirty	where
room	make	right	on
window	ride	left	here
mess	say	ready	in
chute	decide	smart	so
cookie	run	own	away
time	turn	great	down
bread	dry	happy	next
powder	let	pretty	shortly
kitchen	leave	nice	very
floor	watch	whole	again
butter	know	wonderful	how
chocolate	eat	close	off
home	swim	excited	yet

grape	lay	full	shortly	grape	give	fine	over
fishtank	jump	gentle	together	fish	climb	first	sure

60s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
baby	go	little	then
dog	get	good	back
Carl	look	clean	there
bed	put	big	all
back	take	dry	up
crib	have	hungry	just
mother	come	open	now
mom	run	happy	out
laundry	play	great	upstairs
chute	make	next	where
milk	decide	dirty	very
mess	say	left	in
powder	see	nice	so
window	ride	old	down
time	do	sure	probably
bread	find	first	over
floor	watch	messy	shortly
butter	know	pretty	off
cookie	climb	smart	again
chocolate	clean	soft	away
aquarium	let	tired	on
home	dance	able	around
room	sit	close	also

70s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
baby	go	good	then
dog	get	little	back
Carl	look	big	there
back	put	next	all
crib	have	dirty	just
bed	take	open	now
mother	come	nice	out
laundry	see	hungry	up
window	clean	first	again
milk	say	old	upstairs
powder	do	sure	where
floor	know	wonderful	over
chute	turn	happy	very
bread	find	great	off
boy	sit	same	here
mess	make	messy	on
mom	climb	right	in
kitchen	watch	pretty	shortly
cookie	play	cute	so
butter	decide	tired	down
chocolate	dance	whole	next
thing	let	close	away
time	dry	smart	yet

soap	swim	full	here	aquarium	eat	innocent	how
thing	eat	whole	soon	puff	open	own	apparently

80s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
baby	go	little	then
dog	get	good	back
Carl	look	big	there
back	put	next	all
bed	take	happy	up
boy	have	dirty	just
mother	come	nice	now
crib	play	great	out
window	see	open	where
laundry	do	old	on
bread	say	messy	down
milk	clean	smart	again
powder	know	sure	here
floor	watch	right	very
butter	find	ready	upstairs
chute	staighten	wonderful	in
chocolate	make	left	so
puff	ride	small	apparent
child	wash	wet	off
cookie	give	able	over
fish	open	whole	how
head	let	close	shortly
time	climb	first	around

room	guess	pretty	really
thing	run	different	away

Picnic

20s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
mouse	go	little	all
family	get	stuffed	back
picnic	have	happy	there
truck	look	missing	then
baby	start	lost	just
road	eat	red	meanwhile
back	find	sad	out
time	play	pink	where
child	see	good	very
animal	decide	big	still
kid	come	hungry	so
berry	call	left	up
flower	know	bumpy	around
girl	notice	ready	finally
mom	pick	excited	together
rest	do	small	alone
grass	begin	scared	now
day	drive	old	really
rock	cry	same	off
raspberry	run	beautiful	maybe

30s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
mouse	go	little	back
picnic	get	stuffed	all
family	see	missing	there
truck	have	happy	then
road	find	pink	out
baby	look	sad	where
back	eat	good	up
child	start	big	meanwhile
time	realize	lost	just
girl	decide	hungry	so
kid	begin	ready	still
berry	come	bumpy	around
flower	run	whole	very
animal	hear	great	now
doll	know	left	together
raspberry	fall	excited	here
grass	take	scared	finally
food	do	old	behind
baseball	pick	own	alone
rock	continue	young	down

dad	take	own	again	water	sit	small	again
food	hear	whole	behind	toy	give	beautiful	once
baseball	hug	glad	even	lunch	hit	high	even
way	fall	long	here	mom	forget	nearby	really
lake	sit	nice	away	bush	hug	nice	on

40s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
mouse	go	little	all
picnic	look	stuffed	back
truck	get	missing	there
family	have	red	then
road	play	pink	just
baby	eat	happy	where
back	find	lost	out
time	start	good	very
berry	see	big	up
flower	realize	sad	still
rat	come	hungry	so
child	sit	small	around
animal	know	left	here
rock	run	ready	now
food	call	old	together
blanket	cry	beautiful	again
bush	do	scared	finally
doll	take	nice	alone
kid	drive	whole	behind
day	fall	young	off

50s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
mouse	go	little	all
picnic	look	stuffed	back
truck	get	happy	there
family	have	missing	then
road	play	big	where
baby	eat	pink	out
back	see	good	meanwhile
kid	find	sad	just
child	start	lost	so
grandpa	come	hungry	very
mom	sit	ready	up
flower	call	whole	around
raspberry	do	full	still
time	decide	bumpy	now
animal	pick	left	here
berry	realize	beautiful	together
toy	know	wonderful	finally
dad	take	poor	maybe
doll	run	old	behind
blanket	cry	great	down

lake	decide	great	suddenly	grass	drive	small	really
grass	pick	excite	away	food	think	nice	even
toy	lay	bumpy	down	grandma	notice	tall	again
lunch	begin	poor	on	day	hear	lonely	off
area	notice	full	probably	way	continue	scared	alone

60s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
mouse	go	little	all
picnic	look	happy	there
truck	have	good	back
family	get	lost	just
road	play	pink	so
baby	find	big	then
time	eat	missing	out
flower	start	stuffed	very
back	see	red	where
mom	sit	sad	up
doll	come	hungry	around
kid	do	whole	still
berry	take	great	meanwhile
dad	know	ready	maybe
toy	decide	poor	now
grandpa	cry	beautiful	here
child	pick	nice	together
raspberry	run	old	finally
food	fall	young	even

70s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
mouse	go	little	all
picnic	look	good	there
truck	have	happy	back
road	get	big	out
baby	play	lost	then
family	eat	missing	now
time	seem	pink	where
flower	find	red	up
doll	sit	stuffed	so
back	come	nice	still
toy	start	sad	very
berry	take	ready	just
child	pick	old	around
kid	know	poor	here
basket	do	great	meanwhile
water	cry	hungry	again
grandpa	decide	bumpy	maybe
bush	lay	whole	suddenly
tree	run	left	off

brother	realize	bumpy	down	thing	call	full	down
water	drive	left	really	rock	realize	young	meantime
baseball	lay	full	off	rest	fall	beautiful	finally
bush	call	tall	again	place	think	small	even
blanket	swim	wonderful	on	middle	swim	wonderful	together
area	jump	aware	probably	girl	hear	glad	on

80s

Nouns	**Verbs**	**Adjectives**	**Adverbs**
mouse	go	little	there
picnic	have	lost	all
truck	look	big	then
road	get	good	back
baby	play	happy	out
family	eat	ready	just
flower	come	red	where
toy	see	missing	still
time	find	pink	so
back	sit	poor	up
mother	know	sad	here
food	take	old	now
basket	start	bumpy	around
picture	pick	great	very
thing	cry	nice	on
baseball	decide	wonderful	meanwhile
water	think	beautiful	down
berry	do	whole	even

rock	guess	hungry	maybe
place	hug	stuffed	off
watermelon	lay	glad	again
doll	fall	pretty	apparently
home	jump	next	along
banjo	swim	full	away
child	put	small	ever

CHAPTER 4

STUDY II

Function words in narrative discourse in aphasia

ABSTRACT

Purpose

A multitude of measures have been applied to investigate discourse production in persons with aphasia (PWA). However, these measures have not been widely used in clinical settings due to resource-heavy procedures. The purpose of this study was to develop a clinically acceptable measure that evaluates function word production in discourse. To identify the core function word lists, age and narrative task were considered.

Method

Two wordless picture books were used to collect narrative language samples from 470 control speakers (20s, 30s, 40s, 50s, 60s, 70s, and 80s). Core lexicon lists were developed by identifying the 25 most frequently occurring function words for the seven age groups, the two wordless picture books (Good Dog Carl, Picnic), and also combined across both stories. Language samples from 11 PWA were used to determine the relationship among function word production and aphasia determined by the aphasia quotient (AQ) from the Western Aphasia Battery – Revised (Kertesz, 2006).

Results

Significant differences for function word production were found between the two discourse stories, but not across the age cohorts. Because age was not a factor for the function word lists, discourse samples for the PWA were compared to core function lists for each discourse story. Then, Spearman's correlations were performed to determine the relationship among function

word production and aphasia severity (WAB-R AQ). The two core lexicon lists significantly correlated with overall aphasia severity.

Conclusions

These findings have potential clinical and methodological implications. Discourse elicitation stimuli should be considered when elicit heterogeneous usage of function words. Core function word lists may facilitate assessment of language usage at the discourse level by reducing time and resources in clinical settings.

INTRODUCTION

Word retrieval is a common problem for persons with aphasia (PWA) and researchers most commonly have focused on word retrieval difficulties related to retrieval of content words (Bates, Chen, Tzeng, Li, & Opie, 1991; Camarazza & Hills, 1990; Chen & Bates, 1998; Kim & Thompson, 2000; Luzzatti & Chierchia, 2000; Saffran, Schwartz, & Marin, 1980) with relatively few studies examining the retrieval of function words. At the discourse level, function word production (e.g., referents, prepositions) contributes to elaborative phrasing or sentence structure in binding story elements (Halliday & Hasan, 1976). However, function word production at the discourse level has received little attention to date.

Those investigations that have addressed function word production in discourse have reported differences between subtypes of aphasia (e.g., fluent aphasia vs non-fluent aphasia) (Gordon, 2006; Kolk & Heeschen, 1992, 1996; Rochon, Saffran, Berndt, & Schwartz, 2000; Saffran, Berndt, & Schwartz, 1989; Salis & Edwards, 2004). For example, Saffran and colleagues (1989) compared grammatical production between participants with aphasia and agrammatism (N = 5), participants with aphasia and non-agrammatism (N = 5), and cognitively

healthy adults (N = 5). In their study, the participants told the Cinderella story and proportion of closed class words were computed. Closed class words consisted of pronouns, determiners, prepositions, conjunctions, quantifiable adverbs, verb inflections, verb particles, and auxiliary verbs, and they are considered function words. Significant differences for proportion of closed class words were found between the two aphasia groups ($p < .001$) with the agrammatism aphasia group producing a lower proportion of closed class words than the non-agrammatism aphasia group. The non-agrammatism aphasia group and the control group did not significantly differ in proportion of closed class words produced. In a later study, Gordon (2006) attempted to replicate Saffran et al.'s (1989) findings. Gordon included eight participants with fluent aphasia, eight participants with non-fluent aphasia, and six participants without brain damage. Consistent with Saffran and colleagues' findings (1989), the non-fluent aphasia group produced a lower proportion of closed-class words and had less frequent use of obligatory determiners compared to the fluent aphasia group. Gordon concluded that such measures can facilitate clinical judgement in determining agrammatism.

Despite the potential clinical relevance of function word production in PWA, a theoretical explanation of these findings remains obscure. Kolk and Heeschen (1992) proposed function word production at the discourse level requires sufficient processing capacity and PWA (particularly persons with non-fluent aphasia) present with reduced processing capacity; thus they produce simplified utterance structures resulting in omission of function words and referred to as "elliptical style" (Kolk & Heeschen, 1990, p.229). Further, researchers have claimed that task variation affects omission of function words (De Roo, Kolk, & Hofstede, 2003; Indefrey et al., 2001; Kolk, 1995; Salis & Edwards, 2004). Kolk and Heeschen (1996) suggested the adaptation theory which postulates that if a complex task taxes an individual's language

processing capacity, then the PWA attempts to reduce the computational overload by producing simplified utterances; it should be noted that this theory is based on Dutch and German studies but has been applied to interpret findings with English speaking PWA.

Salis and Edwards (2004) replicated Kolk and Heeschen's (1996) findings and included two groups - persons with non-fluent aphasia (N = 4) and a control group (N = 4). They investigated the overuse of elliptical speech and task effect in discourse collected from spontaneous language and picture description tasks. The picture description task consisted of drawings of shapes and colours depicting different spatial positions. Outcome measures included omissions and substitutions of determiners, verbs, inflections, and prepositions. Similar to Kolk and Heeschen's (1996) results, the aphasia group differed from the controls and had more omission of determiners and prepositions. Although Salis and Edwards did not provide further explanation regarding the underlying mechanisms that accounted for these findings beyond that of a limited processing capacity, their main finding was that function word production is associated with grammatical impairment in persons with non-fluent aphasia.

In the current study we focused primarily on developing a function word measure that addresses issues of clinical feasibility. Although only a few studies have examined function word ellipsis in discourse produced by PWA within the same framework raised by Kolk and Heeschen (1996), the notion that omission of function words is a possible strategy to reduce the computational steps to produce connected speech is compelling. Given the need to develop a quantitative measure of function word production for clinical practice, the intent of the study was to examine the presence or absence of function words in discourse produced by PWA. In the following sections, we review the core lexicon research that has led to the computational analysis for this study.

Core Lexicon Measures

Recently, researchers have developed and applied core lexicon analysis to aphasia narratives (Dalton & Richardson, 2015; Kim, Kintz, Zelnosky, & Wright, 2019; Kim & Wright, 2020; MacWhinney, Fromm, Holland, Forbes, & Wright, 2010). Core lexicon measures consist of critical lexical items that play a significant role in constructing a semantically coherent narrative (MacWhinney et al., 2010). The use of core lexicon measures has many advantages. The core lexicon list can be created with computational language analysis programs, such as the Computerized Language Analysis (CLAN, MacWhinney, 2000) program, which reduces analysis errors (Dalton & Richardson, 2015; Dillow, 2011; Fromm, Forbes, Holland, & MacWhinney 2013; MacWhinney et al., 2010). Additionally, core lexicon analysis can be considered clinician friendly because it is not time-intensive to complete or score. Potentially, clinicians could use a checklist of the core lexicon items for a specific narrative discourse task without having to undertake common time-consuming processes that many discourse analysis procedures require (e.g., transcribing, training, and completing the analysis).

MacWhinney and colleagues (2010) included core lexicon measure as a method for demonstrating use of TalkBank tools with the AphasiaBank database. They included language samples from 25 healthy participants to extract the 10 most frequent nouns and verbs by using Computerized Language Analysis (CLAN; MacWhinney, 2000). They found group differences on the core lexicon measure; the aphasia group presented with reduced lexical diversity and greater use of light verbs. Function words were not considered in the study. In a more recent study using a core lexicon measure, Dalton and Richardson (2015) included function words in their core lexicon list. The study included 92 cognitively healthy adults to build core lexicon items, and then they examined if the core lexicon measure could discriminate a different group of

cognitively health participants (N = 166) from PWA (N = 235), as well as, among aphasia subtypes. The cognitively healthy group produced more items on the core lexicon list compared to the aphasia group. Within the aphasia group, participants with fluent aphasia performed better on the measure compared to participants with nonfluent aphasia. The researchers also rated the participants' ability to convey the gist of the story (main concept analysis) to investigate the relationship between core lexicon and main concept performance. Significant correlations between the two measures were found across the subtypes of aphasia (i.e., anomic, Broca's, conduction, Wernicke's). Inclusion of function words as part of the core lexicon list may have contributed to the significant correlations; however, word type was not considered separately (i.e., noun, verb, function word, etc…) to test this hypothesis. Finally, Dalton and Richardson concluded that core lexicon may have clinical utility in understanding word retrieval ability at the discourse level.

In the current study, core lexicon lists of function words are developed to evaluate function word production by PWA. MacWhinney and colleagues' (2010) core lexicon procedures served as the basis for extracting the function word items. Heller and Dobbs (1993) reported age differences in the pattern of function word usage at the discourse level, yet other researchers have not found age-induced changes on lexical or grammatical processing (Kemper, Greiner, Marquis, Prenovost, & Mitzner, 2001; Kemper & Sumner, 2001; Marini et al., 2005). Given the limited and conflicting findings, it is unclear to what extent age plays a role in function word production, and whether it may simultaneously affect the sensitivity of the core lexicon measure to detect aphasia severity. Further, in previous studies, items within core lexicon lists differed across stimuli. The purpose of the current study, then, was to develop a clinically

acceptable measure that evaluates function word production in discourse and considers age and narrative task. The research questions addressed were as follows:

(1) Do core function word lists differ across age groups and narrative elicitation tasks?

(2) Do PWA differ from cognitively healthy participants for functions words produced across the narrative elicitation tasks?

(3) Does percent of core function words from the lists produced by PWA significantly correlate with severity of aphasia determined by the aphasia quotient (AQ) from the Western Aphasia Battery – Revised (WAB-R; Kertesz, 2006)?

METHOD

Participants

The study included language samples from 470 healthy participants (273 females, 197 males) and 11 PWA. The normative data presented are a subset of data from a larger study examining discourse processing across the lifespan (Wright & Capilouto, 2017). The database included discourse samples and cognitive measures collected from 470 participants ranging in age from 20 to 89 years. Control participants were divided into seven age groups (20s, 30s, 40s, 50s, 60s, 70s, and 80s). All control participants (a) were native English speakers, (b) passed hearing (Davis & Silverman, 1978) and vision screenings (Beukelman & Mirenda, 1998), (c) presented with normal cognitive functioning as indicated by the Mini-Mental State Exam (MMSE; Folstein, Folstein, & McHugh, 2001), and (3) self-reported no history of stroke, head injury, or progressive neurogenic disorders. Demographic information for the control participants can be found in Table 4.1.

All PWA met the following criteria: (a) native English speaker, (b) passed hearing and vision screenings, (c) no reported history of other neurological disorders, (d) presented with aphasia as determined by performance on the WAB-R AQ subtests, (e) chronic aphasia (at least 6 months post-onset), and (f) left hemisphere damage. The PWA were recruited from local support groups and university Speech-Language-Hearing clinics. Two participants with aphasia (P2 and P8) were disqualified from the study due to other neurological disorders. Demographic information for the PWA can be found in Table 4.2.

Discourse Elicitation Tasks

Two wordless picture books were used to collect discourse samples from participants. They included *Good Dog Carl* (*GDC*; Day, 1985) and *Picnic* (McCully, 1984). These types of story generation tasks are "more representative of spontaneous communication" (Liles, 1993, cited in Hughes, McGillivray, & Schmidek, 1997, p. 19) compared to procedural tasks or story retelling tasks. Additionally, because participants are telling stories from books rather than from shorter pictured stimuli (e.g., single pictures), participants provided samples with greater lexical diversity and length (Fergadiotis & Wright, 2011; Fergadiotis et al., 2011; Wright, Capilouto, Srinivasan, & Fergadiotis, 2011). *GDC* is a 30-page book that follows a temporally-driven story structure conveying the events that unfold as a dog is left to take care of a baby. *Picnic* is a 31-page story that represents a spatially- and temporally-driven story structure conveying the adventures of a family of mice going on a picnic. For the discourse task, participants were presented with the book and allowed to look through it for as long as they needed. During the story telling phase, participants kept the stimuli in view so that memory was not a factor.

Language Sample Preparation

All samples were either audio or video recorded, and then orthographically transcribed by trained research assistants in CHAT of the Computerized Language Analysis (CLAN; MacWhinney, 2000) program. Inter-rater and intra-rater reliability for word-by-word transcription were determined for 10% of each participant group (i.e., the control group and the aphasia group). For the control group, inter- and intra-rater agreements were 95% and 98%, respectively. For the aphasia group, two persons with aphasia were selected due to the small number of participants, and inter- and intra-rater agreements were 91% and 93%, respectively.

To generate the core lexicon list for function words, we used the GEM, MOR, and FREQ programs associated with CLAN (MacWhinney, 2000). GEM extracts the specific stories under analysis from the larger discourse sample and creates a separate file for each participant containing the specific stories (i.e., GDC and Picnic) under review. To automate the process of finding function words, MOR was used to automatically assign a syntactic category to each word (for review see MacWhinney, 2000; MacWhinney e al., 2010). The MOR program uses a dictionary of lexical items and English grammar rules to assign each word to its respective syntactic category. The MOR program has an accuracy of 95% (for review, MacWhinney et al., 2010). The FREQ program extracts words, word classes, or other coded items from the discourse samples and generates a frequency list. FREQ was utilized to generate a list of all the function words produced within the transcripts along with their frequency information. For this study, the function word lists included pronouns, determiners, prepositions, conjunctions, coordinators, quantifiers, negatives, and copula. The top 25 most frequent function words were used to create the core function word lists for each age group and across both narrative tasks. While the top 25

most frequent function words is an arbitrary cut-off, previous researchers used similar numbers (Dalton & Richardson, 2015).

Core Function Word List

The procedures above were applied to generate 17 core function word lists. List 1 consisted of the most frequent function words across all age groups (age-invariant) and both narratives. Lists 2-3 represented the most frequent function words produced by all age groups (age-invariant) for GDC and Picnic, respectively. Lists 4-10 consisted of the core function words for GDC for each age cohort (20s, 30s, 40s, 50s, 60s, 70s, & 80s), and Lists 11-17 represented the most frequent function words for Picnic for each age cohort. For more information on the different core function word lists, refer to Table 4.3.

Core Function Word Agreement

The core function word lists were utilized to generate percent agreement for the function words produced by the PWA. Computational analyses were used to create the list of function words produced by the PWA. A FREQ code was generated that would automatically count whether a participant produced any of the core function words from the appropriate list (i.e., the matching age and narrative task list), and the number of times the core function word was used. Percent agreement was calculated by giving 1-point for each function word the participant produced that was part of the core function word list, regardless of the number of times the word was produced. For example, if *the, and, a* were part of the core function word list for Story A and Participant 1 produced "*the*" 26 times, "*and*" 12 times, but did not produce "*a*", the participant would receive 2 points out of 3 – one point for "*the*", one point for "*and*," and zero points for "*a*."

RESULTS

To address the first research question of whether the core function word lists differed across age group and narrative task, a mixed measures analysis of variance (ANOVA) was conducted on the percentage of core function words produced from list 1 (age-invariant, combined narratives) by each participant for *GDC* and *Picnic*. Age group was the between-subject variable, and narrative task was the within-subject variable. The rationale for utilizing an age-invariant, combined narrative list is that if said list represents the age cohorts and narrative tasks equally, we would expect performance between the age groups and narrative tasks to be similar and, therefore, not significant. The age group main effect was not statistically significant, but the narrative elicitation task main effect, $F(1, 452) = 9.009$, $p = .003$, $\eta_p^2 = .020$, and age group by narrative elicitation task interaction, $F(6, 452) = 2.616$, $p = .017$, $\eta_p^2 = .034$, were statistically significant. Cognitively healthy adults produced slightly more core function words from List 1 for Picnic (M = 87.84, SD = 8.18) compared to GDC (M = 87.76, SD = 8.26). The results indicate that a core function word list needs to be narrative specific but not age specific. The non-significant age main effect suggests that the age corrected lists (4-17) can be removed from further analysis. The significant narrative main effect indicates that List 1 can be removed from further analysis. The remaining analyses utilized List 2 (age-invariant, *GDC*) and List 3 (age-invariant, *Picnic*). See Appendix 4.A for the core lexicon list for age-invariant and combined narratives of both narratives.

To address the second research questions of whether PWA differ in core function word production compared to cognitively healthy adults, an agreement analysis was conducted by comparing the percentage of core function words produced by PWA with List 2 (age-invariant, *GDC*) and List 3 (age-invariant, *Picnic*). For GDC, PWA's agreement ranged from 20% to 88%

(M = 53.60%, SD = 26.41%). For Picnic, PWA's agreement ranged from 8% to 80% (M = 42.54%, SD = 27.96%). These results indicate that PWA are producing fewer and/or a different variety of function words when compared cognitively healthy adults. Complete lists (Lists 2 & 3) of the top 25 core lexicon are presented in Appendix 4.B. Table 4.4 presents the percent agreement PWA and cognitively healthy adults for core function word lists for GDC and Picnic.

---------- Table 4.4 about here----------

To address the third research question of whether the percentage of core function words produced by PWA significantly correlates with aphasia severity, a bivariate Spearman's correlation was conducted on the percentage of function words produced for Lists 2 and 3 (age-invariant, story specific) with the participant's AQ. Following Goodwin and Leech (2006)'s guidelines, Spearman's correlations were used because: (1) visual inspection of the histograms, as well as Q-Q plots, determined a lack of normality, (2) large standard deviations suggested larger variability potentially inflates r-scores, and (3) the small n ($n < 30$) can exacerbate any problems associated with non-normality and larger variability. Significant correlations for *GDC* and *Picnic* were found between percent agreement of function word production and overall aphasia severity determined by WAB-R AQs, $r = .825$, $p < .00$ and $r = .589$, $p < .001$, respectively.

DISCUSSION

The goal of this study was to develop core lexicon lists of function words and examine if function word production by PWA using the core function word lists correlated with aphasia severity. Age-related differences were not found for core function words; however, stimuli-related differences were. As such, moving forward with determining the relationship among

function word production and aphasia severity, age-invariant core function word lists for each stimulus were used. Results indicated that function word production using these core function word lists significantly correlated with aphasia severity. As follows are discussion of the results and potential clinical implications.

Core Function Words and Age

Based on previous findings with core lexicon lists for other word types (i.e., nouns and verbs), we hypothesized that age would play a significant role in core function word production. The nonsignificant findings were initially surprising and differed from previous findings reporting age-related differences on lexical retrieval (Dennis & Hess, 2016; Fergadiotis et al., 2011; Kave & Goral, 2017; Marni et al., 2005; Shewan & Henderson, 1998). One possible explanation for our finding may be that fewer function words compared to other word types are available in our daily life, thus increasing the likelihood of producing the same function words. Baayen, Piepenbrock, and Gulikers (1995) reported that the average English speaker has a productive vocabulary of more than 100,000 words. However, function words account for less than 0.04% of this total. Therefore, the lack of age-related differences on the lists is not as surprising; developing and applying an age-invariant list to quantify function words in discourse may be appropriate.

The current findings demonstrated that cognitively healthy adults produced different core function words from the age-invariant and combined core lexicon list (List 1) for the two narrative tasks (*Good Dog Carl* and *Picnic*). Though the means of percent agreement for List 1 with each of the narratives appear very similar (i.e., 87.84% v. 87.76%), since the difference was statistically significant, we conservatively interpreted these findings and used the age-invariant, narrative-specific core function word lists (i.e., Lists 2 and 3) for

subsequent analyses. The discourse stimuli present with different story structures, relationships among characters, and sequencing of events which have been reported in previous research (Wright & Capilouto, 2012; Wright, Capilouto, Srinivasan, & Fergadiotis, 2011). When language samples collected from different stories are compiled for a core lexicon list, the inherent properties of each of the stories may be dampened. For example, the pronoun *them* is included in the core lexicon list for *Picnic* and not in that of *Good Dog Carl* and the combined story. *Them* is a unique word that may be produced by speakers when building their story for the *Picnic* book; because, the opportunity to produce the word is available in *Picnic*, but not in *Good Dog Carl*. Another example is the conjunction *but* – it was one of the 25 most frequently occurring function words for *Good Dog Carl,* but not for the other two lists (*Picnic*, combined story). The difference observed in preposition use is not as easily and clearly explained. It may be associated with story structure and story setting (i.e., characters, story context). To tell the *Good Dog Carl* story, each event depicted occurs in a different location (e.g., the bedroom, living room, kitchen, and bathroom) and should be described. The events/actions continue as the story progresses in a temporal fashion. In language samples elicited based on *Good Dog Carl*, *but* occurs most often when speakers describe the sequence of events (e.g., the baby made a mess *but* didn't stop, they played with everything *but* that's not enough). Evaluating accuracy for use of the function words produced was not determined and was beyond the scope of the current study. However, it may be considered in future studies of function word production.

Function Word Production and Aphasia

As suggested by Webster and colleagues (2007), the performance of cognitively healthy speakers provides a baseline to examine degree of language impairment in clinical populations. Surprisingly, few studies have considered cognitively-healthy standards for comparisons

(MacWhinney et al., 2010). The current study developed core function word lists from normative data to investigate discourse-level function word production by PWA. The range for PWA's percent agreement score was large, indicating that the variation of percent agreement scores may have potential to capture a range of different language abilities among our participants with aphasia. Statistically significant correlations were found among WAB-R AQs and percent agreement for function words produced using the core function word lists for both stories (i.e., *GDC* and *Picnic*), indicating fewer function words produced as severity of aphasia increased. Our findings support Cui and Zhong's (2018) assertion that it is possible that the presence and absence of function words are related to the degree of aphasia severity. They suggested that reduced function word production can be explained within Kolk and Heeschen's (1996) framework. PWA tend to use an elliptical strategy to compensate for their reduced cognitive sources and resulting in reduced discourse production output (Cui & Zhong, 2018). Possibly then, performance of function words can improve as recovery progresses and should be considered in future treatment and longitudinal studies.

Although it is well known that function word processing is primarily involved in syntactic processes (Bradley, Garret, & Zurif, 1980; Gordon & Caramazza, 1982; Kolk & Blomert, 1985; Shillcock & Bard, 1993), both semantic and syntactic processes are activated in tasks beyond the single word level (Cole & Segui, 1994; Friederici, 1982, 1985; Friederici, Opitz, & Cramon, 2000). Semantic and syntactic information come into play when prepositions and pronouns are produced. Grodzinsky (1984) suggested that pronouns might induce a concrete image for content words. For example, the pronoun *she* indicates a singular female image, requiring an active, semantic representation of the antecedent. Reflecting a similar viewpoint, Bird, Franklin, and Howard (2002) demonstrated that function words with a concrete image were

produced more frequently in their aphasia participants' spontaneous speech (N = 5). For example, gender category (e.g., *he, him, his*) resulted in better performance for their aphasia participants compared to other categories of function words in a semantic judgement task. Additionally, Friederici (1982) suggested that prepositions which refer to location (i.e., in, out, up, down) rely on successful access of semantic and syntactic processing to produce and comprehend referential meaning, particularly for persons with Broca's aphasia. Appendix B presents the 25 most frequently occurring function words in the cognitively healthy control group. Pronouns and prepositions account for a substantial proportion of the core function word listed. Collectively, function words that rely on both semantic and syntactic contributions for production constituted the core function word lists in the current study and this may contribute to why significant correlations were found with the measure of aphasia severity.

Core function word production for aphasia subtypes was not statistically investigated in the study due to the small number of PWA; however, visual inspection of the data indicates that the six PWA who presented with a non-fluent type of aphasia had the lowest percent agreement for the function words produced, regardless of narrative task. In previous studies, researchers have reported differences in function word production between participants with fluent and non-fluent types of aphasia (e.g., Friederici, 1982, 1985; Friederici, Opitz, & Cramon, 2000). Gordon (2006; 2008) and Saffran, Berndt, and Schwartz, (1989) reported similar patterns for persons with fluent and non-fluent aphasia during connected speech – participants with non-fluent aphasia produced fewer closed class words and use of determiners compared to participants with fluent aphasia. Gordon (2008) interpreted her findings as evidence of a trade-off between semantic and syntactic processes occurred in fluent and non-fluent aphasia in opposite directions. More specifically, PWA with relatively intact syntactic ability (i.e., persons with fluent aphasia)

tended to show reduced semantic ability, indicating that these individuals produced a higher proportion of syntactically laden words, such as function words, in connected speech. PWA with decreased syntactic ability (i.e., persons with non-fluent aphasia) tended to rely on semantic processing, which led to the production of a higher proportions of semantically laden words and a lower proportion of syntactically laden words. Earlier studies of function words (e.g., Friederici, 1982, 1985; Friederici, et al., 2000) have suggested that there are more semantically- and syntactically- related words within the category of function words. Together, by creating subsets of core function words depending on the relative weights of syntactic and semantic processing, future studies could investigate the utility of core function word production for quantifying differences among aphasia subtypes, which will help us broaden our understanding of function word processing in PWA's connected speech.

Measurement Issues

The core function word measure provides an advantage to other methods quantifying function word production because it reduces the work-load burden on clinicians. Scoring can be completed without orthographically transcribing samples. However, it can be argued that the use of core function word lists does not allow for identification of error production of function words in PWA language samples. As noted earlier, the study of Dalton and Richardson (2015) demonstrated that counting the presence of lexical items is a valid means of quantifying word retrieval ability at the discourse level. Moreover, our approach was motivated by the idea of the adaptation strategy that a handful of studies have investigated regarding the production of simplified utterances (i.e., omission of function words) in PWA's discourse, which supports this way of scoring (e.g., De Roo, Kolk, & Hofstede, 2003; Kolk, 1995; Kolk & Heeschen, 1992; Ruiter, Kolk, Reitveld, 2010; Ruiter, Kolk, Rietveld, & Feddema, 2013; Salis & Edwards, 2004).

The concept of the adaptation strategy in which the omission of function words in PWA's discourse could be a manifestation of a reduced linguistic and/or cognitive capacity has been theoretically substantiated. Considering the limited time and resources in clinical settings, defining a grammatical category for each word in language samples in traditional ways may deter use of discourse analyses. Identifying the error production and grammatical category of function words in discourse produced by PWA is admittedly important in quantifying function word production, it may come at the expense of clinical utility of the measure. Development of the core lexicon measure as a clinician-friendly method for discourse analysis is still in its early stage. Subsequent studies are needed to determine reliability and validity of the measure.

Since core lexicon analysis is a relatively new measure, no consensus exists among researchers for defining the pre-determined core lexicon items. MacWhinney and colleagues (2012) generated 10 core nouns and 10 core verbs from a discourse task (approximately 20% of all lexicon items produced). Fromm, Forbes, Holland, and MacWhinney (2013) identified 10 nouns and 10 verbs based on how frequently the lexical items occurs by the control participants. Dalton and Richardson (2015) aggregated all word classes in a core lexicon list with lexical items produced by greater than 50% of the sampling cohort. Only the core lexicon list developed by Dalton and Richardson included function words. In developing a new language test, including items of varying difficulty enhances the sensitivity of indexing language impairments (Ivanova & Hallowell, 2013). Further, while there is a precedent for defining text from a frequency list (Gottron, 2009), the cut-off we used (25 most frequently produced function words) was arbitrarily determined with ease of use being the most important factor in that decision. Thus, the number of items in our core lexicon list was similar to the numbers identified in previous studies

(Dalton & Richardson, 2015). Future studies should consider a systematic approach to the establishing criterion for determining core lexical list length and items included.

CONCLUSIONS AND FUTURE DIRECTIONS

Results of this study extend previous findings by researchers who have investigated function word production in PWA. We did not find significant age-related for function word production. For PWA, percent of function words produced using the core function word lists significantly correlated with aphasia severity.

Findings from the study have important clinical and methodological implications. First, based on the non-significant differences among age-cohorts, our results suggest that age factors do not need to be considered in future studies investigating function word production at the discourse level. Similar to previous studies, we did not discriminate sub-categories of function words for scoring (Dalton & Richardson, 2015). The benefit of this scoring system is that it provides a time-efficient tool that may be have potential clinical utility for clinicians. Future investigations are needed that include a larger sample of participants with aphasia to strengthen conclusions regarding the clinical feasibility.

In future studies with clinical populations, researchers could develop core function word lists by sub-categories (e.g., determiners, prepositions, pronouns) (De Roo et al., 2003; J. K. Gordon, 2006; Kemmerer, 2005; Ruigendijk & Bastiaanse, 2002; Salis & Edwards, 2004). For example, De Roo and colleagues (2003) demonstrated pronoun use in discourse differs between cognitively healthy individuals and PWA. Additionally, production of function words contributes to elaborated phrase and sentence structures and coherent discourse (Halliday & Hasan, 1976). Though not examined in this study, performance of core function words may be related to

cohesion and/or coherence in discourse and should be considered in future investigations. Finally, given that PWA experience working memory and attention deficits in communicative activities, investigating the relationship among cognitive functions and function word production at the discourse level would help broaden our understanding of cognitive constructs that influence daily communication.

Acknowledgements

This research was partially supported by National Institute on Aging Grant R01AG029476. We are especially grateful to the study participants.

References

Bates, E., Chen, S., Tzeng, O., Li, P., & Opie, M. (1991). The noun-verb problem in Chinese aphasia. *Brain and language, 41*(2), 203-233.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Beukelman, D. & Mirenda, P. (1998). *Augmentative and Alternative Communication: Management of severe communication disorders in children and adults* (2nd Ed.). Baltimore, MD: Paul H. Brooks Publishing.

Bird, H., Franklin, S., & Howard, D. (2002). 'Little words'—not really: Function and content words in normal and aphasic speech. *Journal of Neurolinguistics, 15*(3), 209-237.

Bradley, D. C., Garrett, M. E., & Zurif, E. B. (1980). Syntactic deficits in Broca's aphasia. In. D. Caplan (ed). *Biological Studies of Mental Processes*. Cambridge: MIT Press.

Caramazza, A., & Hillis, A. E. (1990). Where do semantic errors come from?. *Cortex*, *26*(1), 95-122.

Chen, S. & Bates, E. (1998). The dissociation between nouns and verbs in Broca's and Wernicke's aphasia: Findings from Chinese. *Aphasiology, 12*(1), 5-36.

Cole, P., & Segui, J. (1994). Grammatical incongruency and vocabulary types. *Memory & Cognition, 22*(4), 387-394.

Cui, G., & Zhong, X. (2018). Adaptation in aphasia: revisiting language evidence. *Aphasiology*, *32*(8), 855–875. https://doi.org/10.1080/02687038.2018.1458068

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*, *39*(11), 1125–1137. https://doi.org/10.1044/2015_AJSLP-14-0161

Davis, H. & Silverman, S.R. (Eds.). (1978). *Hearing and deafness* (4th Ed.). New York, NY: Holt, Rinehart, & Winston.

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

Dennis, P. A., & Hess, T. M. (2016). Aging-related gains and losses associated with word production in connected speech. *Aging, Neuropsychology, and Cognition*, *23*(6), 638-650.

De Roo, E., Kolk, H., & Hofstede, B. (2003). Structural properties of syntactically reduced speech: A comparison of normal speakers and Broca's aphasics. *Brain and Language*, *86*(1), 99–115. https://doi.org/10.1016/S0093-934X(02)00538-2

Dillow, E. (2011). *Narrative Discourse in Aphasia: Main Concept and Core Lexicon Analyses of the Cinderella Story*. Columbia: University of South Carolina.

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.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (2001). *Mini mental state examination*. Lutz, FL: Psychological Assessment Resources, Inc.

Friederici, A. D. (1982). Syntactic and semantic processes in aphasic deficits: The availability of prepositions. *Brain and Language, 15*(2), 249-258.

Friederici, A. D. (1985). Levels of processing and vocabulary types: Evidence from on-line comprehension in normals and agrammatics. *Cognition, 19*(2), 133-166.

Friederici, A. D., Opitz, B., & Von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. *Cerebral Cortex, 10*(7), 698-705.

Fromm, D. A., Forbes, M., Holland, A., & MacWhinney, B. (2013). PWAs and PBJs: Language for describing a simple procedure.

Gordon, J. K. (2006). A quantitative production analysis of picture description. *Aphasiology*, *20*(02–04), 188–204. https://doi.org/10.1080/02687030500472777

Gordon, J. K. (2008). Measuring the lexical semantics of picture description in aphasia. *Aphasiology*, *22*(7-8), 839-852.

Gordon, B., & Caramazza, A. (1982). Lexical decision for open-and closed-class words: Failure to replicate differential frequency sensitivity. *Brain and Language, 15*(1), 143-160.

Gottron, T. (2009). Document word clouds: Visualising web documents as tag clouds to aid users in relevance decisions. In *International Conference on Theory and Practice of Digital Libraries* (pp. 94–105). Springer.

Grodzinsky, Y. (1984). The syntactic characterization of agrammatism. *Cognition*, *16*(2), 99-120.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. *Longman, London*.

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], 2017-03-02. http://doi.org/10.3886/ICPSR36634.v1

Heller, R. B., & Dobbs, A. R. (1993). Age differences in word finding in discourse and nondiscourse situations. *Psychology and Aging, 8*(3), 443.

Hughes, D. L., McGillivray, L., & Schmidek, M. (1997). *Guide to narrative language: Procedures for assessment*. Eau Claire, WI: Thinking Publications.

Indefrey, P., Brown, C. M., Hellwig, F., Amunts, K., Herzog, H., Seitz, R. J., & Hagoort, P. (2001). A neural correlate of syntactic encoding during speech production. *Proceedings of the National Academy of Sciences*, *98*(10), 5933–5936. https://doi.org/10.1073/pnas.101118098

Ivanova, M. V., & Hallowell, B. (2013). A tutorial on aphasia test development in any language: Key substantive and psychometric considerations. *Aphasiology*, *27*(8), 891–920. https://doi.org/10.1080/02687038.2013.805728

Kavé, G., & Goral, M. (2017). Do age-related word retrieval difficulties appear (or disappear) in connected speech? *Aging, Neuropsychology, and Cognition*, *24*(5), 508-527.

Kemmerer, D. (2005). The spatial and temporal meanings of English prepositions can be independently impaired. *Neuropsychologia*, *43*(5), 797–806.

Kemper, S., & Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and aging*, *16*(2), 312.

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K., & Mitzner, T. L. (2001). Language decline across the life span: Findings from the nun study. *Psychology and aging*, *16*(2), 227.

Kertesz, A. (2006). *The Western aphasia battery-Revised*. San Antonio, TX: Pearson.

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,* 62–78.

Kim, M., & Thompson, C. K. (2000). Patterns of comprehension and production of nouns and verbs in agrammatism: Implications for lexical organization. *Brain and Language, 74*(1), 1-25.

Kim, H. & Wright, H.H. (2020). Concurrent Validity and Reliability of the Core Lexicon Measure as a Measure of Word Retrieval Ability in Aphasia Narratives. *American Journal of Speech-Language Pathology, 29*(10), 101-110.

Kolk, H. (1995). A Time-Based Approach to Agrammatic Production. *Brain and Language, 50*(3), 282–303. https://doi.org/10.1006/brln.1995.1049

Kolk, Herman, & Heeschen, C. (1990). Adaptation Symptoms and Impairment Symptoms in Broca's Aphasia. *Aphasiology*, *4*, 221–231. https://doi.org/10.1080/02687039008249075

Kolk, H., & Heeschen, C. (1992). Agrammatism, paragrammatism and the management of language. *Language and Cognitive Processes*, *7*(2), 89-129.

Kolk, Herman, & 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

Liles, B. Z. (1993). Narrative discourse in children with language disorders and children with normal language: A critical review of the literature. *Journal of Speech, Language, and Hearing Research, 36*(5), 868-882.

Luzzatti, C., & Chierchia, G. (2002). On the nature of selective deficits involving nouns and verbs. *Italian Journal of Linguistics, 14*, 43-72.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database*. MIT Press.

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., & Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiology*, *24*(6–8), 856–868. https://doi.org/10.1080/02687030903452632

Marini, A., Boewe, A., Caltagirone, C., & Carlomagno, S. (2005). Age-related differences in the production of textual descriptions. *Journal of Psycholinguistic Research, 34*(5), 439-463.

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

Rochon, E., Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language*, *72*(3), 193-218.

Ruigendijk, E., & Bastiaanse, R. (2002). Two characteristics of agrammatic speech: Omission of verbs and omission of determiners, is there a relation? *Aphasiology*, *16*(4–6), 383–395. https://doi.org/10.1080/02687030244000310

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and language, 37*(3), 440-479.

Saffran, E. M., Schwartz, M. F., & Marin, O. S. (1980). The word order problem in agrammatism: II. Production. *Brain and language*, *10*(2), 263-280.

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

Shewan, C. M., & Henderson, V. L. (1988). Analysis of spontaneous language in the older normal population. *Journal of Communication Disorders*, *21*(2), 139-154.

Shillcock, R. C., & Bard, E. G. (1993). Modularity and the processing of closed-class words. In Cognitive models of speech processing, *The second sperlonga meeting* (pp. 163-185). East Sussex: Lawrence Erlbaum Associates.

Webster, J., Franklin, S., & Howard, D. (2007). An analysis of thematic and phrasal structure in people with aphasia: What more can we learn from the story of Cinderella? *Journal of Neurolinguistics*, *20*(5), 363–394.

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

Wright, Heather Harris, 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)

Table 4.1

Neurologically Healthy Adult Demographic Information

Age Group	N (F:M)	Age (SD)	Education (SD)
20s	66 (35:31)	24.3 (2.7)	15.8 (2.0)
30s	63 (39:24)	34.1 (3.2)	16.0 (3.2)
40s	67 (41:26)	44.4 (3.1)	15.4 (2.5)
50s	68 (43:25)	53.5 (2.6)	15.8 (2.5)
60s	67 (38:29)	65.2 (4.5)	15.6 (2.8)
70s	76 (43:33)	73.5 (2.9)	15.4 (2.3)
80s	63 (34:29)	83.3 (2.5)	15.2 (3.0)
Total	470 (273:197)	54.4 (19.9)	15.6 (2.6)

Table 4.2

Participants with Aphasia Demographic Information

	Age	Gender	Education	WAB-R AQ[1]	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
Mean	61.7		14.1	68.8	
SD	14.7		2.7	14.0	

[1]Western Aphasia Battery-Revised Aphasia Quotient (Kertesz, 2006)

Table 4.3

Core Function Word Lists

List	Narrative	Age Cohort
List 1	Combined[1]	All
List 2	GDC[2]	All
List 3	Picnic	All
List 4-10	GDC	Age-Corrected
List 11-17	Picnic	Age-Corrected

[1]Combined: *GDC & Picnic*; Age-Correct: Separated by Age Cohort (20s, 30s, 40s, 50s, 60s, 70s, 80s); [2]Good Dog Carl

Table 4.4

Percent agreement between PWA and cognitively healthy adults for core function word lists for Good Dog Carl (GDC) and Picnic

ID	AQ[1]	List 2: *GDC*	List 3: *Picnic*
01A	76.3	64	60
03A	85.2	88	80
04A	62.6	84	68
05A	57.6	24	28
06A	56.3	NA	16
07A	90.7	72	64
09A	57.7	24	12
10A	61.3	48	24
11A	64.9	36	8
12A	89.4	76	80
13A	54.4	20	28
Mean	68.7	53.6	42.5
SD	13.9	26.4	27.9

[1]Western Aphasia Battery-Revised Aphasia Quotient (Kertsz, 2006)

Appendix 4.A: The core lexicon list for age-invariant and combined narratives of both stories – List 1

the	and	be	they	to
a	he	of	in	On
she	up	out	all	that
her	with	him	look	some
For	not	into	I	so

Appendix 4.B: The top 25 core lexicon produced by the control group for the Lists 2 and 3

Age-invariant core lexicon items for Good Dog Carl (List 2)				
the	and	a	he	they
in	on	be	of	to
him	into	some	with	so
she	look	I	up	for
not	out	at	that	but

Age-invariant core lexicon items for Picnic (List 3)				
the	and	they	a	be
he	of	in	she	to
on	so	for	some	with
him	look	not	into	her
all	at	up	I	them

CHAPTER 5

STUDY III

Concurrent validity and reliability of the core lexicon measure as a measure of word retrieval ability in aphasia narratives

ABSTRACT

Background

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 (PWA), the psychometric properties of core lexicon measures have yet to be investigated.

Aim

The purpose of this study was (1) to examine the concurrent validity by using micro- and macro-linguistic measures and (2) to demonstrate inter-rater reliability without transcription by raters with minimal training.

Method

Eleven language samples collected from PWA were used in this study. Concurrent validity was assessed by correlating performance on the core lexicon measure with micro- and macro-linguistic measures. For inter-rater reliability, four raters who had not previously used the core lexicon checklist scored audio-recorded discourse samples of ten PWA.

Results

The core lexicon measures significantly correlated with micro- and macro linguistic measures. Acceptable inter-rater reliability was obtained among the four raters.

Conclusions

Core lexicon analysis is potentially useful for measuring word retrieval impairments at the discourse level. It may also be a clinically feasible solution because it reduces the amount of preparatory work for discourse assessment.

INTRODUCTION

Discourse deficits that negatively impact daily communication for persons with aphasia (PWA) are well known. As such, a variety of approaches for evaluating the meaningful changes in discourse for PWA 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, Ferguson, & Spencer, 2016; 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 patient's 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 PWA such as correct information unit (CIU; Nicholas & Brookshire, 1993) and main concept analysis (Nicholas & Brookshire, 1995). A multi-level 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 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 paper, 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 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 PWA. The authors reported that core lexicon analysis is one method to contrast patterns of lexical usage in PWA in comparison to normal expectations. The core lexicon refers to the pivotal lexical items required to produce a semantically meaningful and coherent narrative. MacWhinney and colleagues (2010) used discourse samples of the Cinderella story from cognitively healthy participants (N = 25) in the AphasiaBank database (MacWhinney, 2000). The ten most frequently occurring nouns and verbs were identified as core lexicon items. Then, they examined whether PWA (N = 24) produced these target words to convey the Cinderella story. The PWA demonstrated a reduced number of core lexicon items and greater use of light verbs (i.e. semantically unspecified verbs such as *be*,

have, take etc…). 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 PWA (N = 92) and control participants (N = 166). The researchers also used main-concept analysis (MC), 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 PWA's 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*, 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 PWA 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 non-fluent 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 pre-determined lexical items. Instead of arduous transcription processes; potentially, clinicians can check if the pre-determined 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 PWA from cognitively healthy controls (Dalton & Richardson, 2015; MacWhinney et al., 2010) and among aphasia subtypes (Dillow, 2011; 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 and colleagues (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 two-fold: (1) to examine the relationship among core lexicon measures and other linguistic measures (micro- and macro-linguistic measures) and (2) to determine inter-rater

reliability for the core lexicon measure using procedures and raters with minimal training. The micro-linguistic measures included syntactic complexity, percent of information units produced, and lexical diversity using the moving average type token ratio (MATTR) method (Covington, 2007; Covington & McFall, 2010). The macro-linguistic 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 micro-linguistic 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 micro-linguistic 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 macro-linguistic 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 (6 female, 5 male) were used in the study, which were also included in Kim and colleagues (2019). The participants' mean age was 61.7 (SD = 14.7) years old 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 25dB HL for the frequencies of 500 Hz, 1000 Hz, 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 PWA can be found in Table 5.1.

Narrative Discourse Task

Two wordless picture books were used to collect narrative discourse samples from participants. They included *Good Dog Carl* (*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.

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 (CLAN; MacWhinney, 2000). In order to meet the aims of the study, micro-linguistic (syntactic complexity, information units, lexical diversity) and macro-linguistic (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).

Inter-rater and intra-rater reliability for word-by-word transcription were measured for two PWA who were randomly selected. Inter- and intra-rater transcription agreements were 91% and 93%, respectively. For c-unit segmenting, inter-rater agreement was 83% and intra-rater 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 60's was evaluated by using the 60's age group core lexicon lists. Synonyms were not counted due to the importance of producing the specific target words (e.g., Andreetta et al., 2012; Verhaegen & Poncelet, 2013). Again, if a PWA produced any target lexical items, they would receive one point. Regardless of how many times the target word may have been used by the participant, only one 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, i.e., if a participant produces 5 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 then CI was computed by adding the total number of independent and dependent clauses and dividing by the number of independent clauses. Inter- and intra-rater agreement for calculating CI was completed for 20% of the transcribed samples. All agreements were above 90%.

Information Units

An Information Unit (IU) is operationally defined as a word that is intelligible, relevant, accurate, and informative relative to the given stimulus. Information units were determined based on previously developed guidelines (Dijkstra, Bourgeois, Allen, & Burgio, 2004; Nicholas & Brookshire, 1993). To determine the percentage of information units (%IU) produced, the number of information units was divided by the total number of words produced and then multiplied by 100. Words included all intelligible word, regardless of their relevance, accuracy, and/or informativeness relative to the stimulus. Inter- and intra-rater 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, Moving Average Type Token Ratio (MATTR; Covington, 2007; Covington & McFall, 2010). MATTR calculates lexical diversity using a moving window to estimate TTRs for consecutive non-overlapping 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 CLAN (Fergadiotis & Wright, 2011) 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), 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 completing a training protocol for scoring coherence. The protocol included language samples to practice scoring and review for accuracy in scoring. Intra- and inter-rater 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 intra- and inter-rater reliability were above 90%.

Rater Reliability

Of all PWA, 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 three to four 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 participant's 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 size (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 information units 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 5.2).

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 information units and lexical diversity, $r = .636$; $p < .05$ and $r = .701$, $p < .05$, respectively. No significant correlations were found among core adverbs and the micro- and macro-linguistic measures (See Table 5.3).

Reliability Analyses

To determine reliability coefficients, intra-class correlation coefficients (ICC; 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., Chicago, Illinois). 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. Table 5.4 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 are required. With respect to concurrent validity, performance of core lexicon production correlated with both micro- and macro-linguistic 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 PWA. 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 PWA's language samples. Therefore, it should be noted that the results present preliminary validity and reliability data on the core lexicon measure.

Core Lexicon and Micro-linguistic Measures

It was hypothesized that the performance of core lexicon production would significantly correlate with micro-linguistic measures (information units, syntactic complexity, and lexical diversity). We found 11 statistically significant correlations among the core lexicon measures and micro-linguistic 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., 2012). 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; PWA 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 PWA's 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 PWA's 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 PWA's 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 on-line.

Before moving forward to the other findings, it is important to recognize structural differences of the two stories implemented in the current study (See S1). The two stories present with different story structure formats such as settings and problems (Wright & Capilouto, 2012; Wright, Capilouto, Srinivasan, & Fergadiotis, 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 and colleagues (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. S2 includes two story examples produced by a participant with aphasia.

PWA's informativeness positively correlated with greater use of core adjectives for *Picnic*, but not for *GDC.* For the *Picnic* story, the percent IUs conveyed increased as PWA produced more core adjectives. This finding is of particular interest, taken together with Wright

and colleagues' study (2011), 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 speakers' 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 PWA. Sarno, Postman, Cho, and Norman, (2005) also suggested that production of adjectives manifested qualitative changes in PWA's 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. PWA 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 pre-determined 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 PWA's 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, Cantagallo, & Marini, 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 micro-linguistic 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 meaning 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 PWA's 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 PWA's 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 Macro-Linguistic Measures

Supporting and extending previous research, several significant correlations emerged among the core lexicon measures and macro-linguistic 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 was 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, Cantagallo, & Marini, 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 macro-linguistic measures for both stories. As PWA 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 inter-rater 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., non-transcription), 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 standard error of measurement (SEM) values for some of the variables were higher than expected. Large SEM values are not ideal for an assessment because this would indicate large measurement error. A practical point to note is that in the usual calculation of SEM, the standard deviation of the PWA' scores are 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 PWA 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 inter-rater reliability for core lexicon analysis. Core lexicon performance by PWA significantly correlated with micro- and macro-linguistic 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 PWA, demonstrating that clinically acceptable inter-rater 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 non-fluent 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 non-fluent 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 inter-relationship among core verb retrieval and micro- and macro- linguistic 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 and colleagues (2005) found that production of modifiers manifested qualitative changes in PWA's

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 sub-categorizations of adverbs (e.g., locative adverbs, prepositional adverbs, and quasi-nominal adverbs) have been considered to play a distinct role or characteristic in utterances (Gilquin, 2007; Pérez-Paredes, Hernández, & Jiménez, 2011). Therefore, research pertaining to PWA's modifiers 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 micro-linguistic 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.

Acknowledgements

This research was partially supported by National Institute on Aging Grant R01AG029476. 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

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

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

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.

Bryant, L., Spencer, E., & Ferguson, A. (2016). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology*, 1-22.

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.

Coelho, C. A. (2002). Story narratives of adults with closed head injury and non-brain-injured adults: Influence of socioeconomic status, elicitation task, and executive functioning. *Journal of Speech, Language, and Hearing Research, 45*, 1232-1248.

Covington, M. A. (2007). *MATTR user manual*. 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.

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

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

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

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. (2011). *Narrative discourse in aphasia: Main concept and core lexicon analyses of the Cinderella Story* (Master's Thesis, University of South Carolina, Columbia). University of South Carolina. Retrieved from Proquest. (UMI #: 1542701).

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.

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 : JSLHR*, *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)

Fromm, D. A., Forbes, M., Holland, A., & MacWhinney, B. (2013, May-June). *PWAs and PBJs: Language for describing a simple procedure.* Poster presented at the Clinical Aphasiology Conference, Tucson, AZ.

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. (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

Guo, Y. E., Togher, L., & Power, E. (2014). Speech pathology services for people with aphasia: What is the current practice in Singapore? *Disability and Rehabilitation*, *36*(8), 691–704. https://doi.org/10.3109/09638288.2013.804597

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

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

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], 2017-03-02. http://doi.org/10.3886/ICPSR36634.v1

Harvill, L. M. (1991). Standard Error of Measurement: an NCME Instructional Module on. *Educational Measurement: Issues and Practice*, *10*(2), 33–41.

Kertesz, A. (2006). *The Western aphasia battery-Revised*. San Antonio, TX: Pearson.

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

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.

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* (Vol. 18). 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 CHILDES project: The database Psychology Press.

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.

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

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.

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.

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

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, Language, and Hearing Research*, *38*(1), 145-156.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory (McGraw-Hill Series in Psychology)* (Vol. 3). McGraw-Hill New York.

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

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

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

Rogalski, Y., Altmann, L. J. P., 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. 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), pp. 83-107.

Schneider, P., Dubé, R. V., & Hayward, D. (2005). The Edmonton narrative norms instrument. *Retrieved from University of Alberta Faculty of Rehabilitation Medicine website: http://www. rehabresearch. ualberta. ca/enni (27.2. 2015)*.

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.

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.

Verna, A., Davidson, B., & Rose, T. (2009). Speech-language pathology services for people with aphasia: A survey of current practice in Australia. *International Journal of Speech-Language Pathology*, *11*(3), 191–205. https://doi.org/10.1080/17549500902726059

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 standardisation 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 multi-level approach to understanding maintenance of global coherence in adults with aphasia. *Aphasiology*, *26*(5), 656-672.

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. 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.

Table 5.1

Participants with Aphasia Demographic Information

	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
P11	72	M	12	64.9	Transcortical motor
P12	65	F	11	89.4	Anomic
P13	65	M	14	54.4	Broca's

Table 5.2

Correlation coefficients (r) among the core lexicon lists and linguistic measures for Good Dog Carl

	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

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

Table 5.3

Correlation coefficients (r) among the core lexicon lists and linguistic measures for Picnic

	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***

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

Table 5.4

Inter-rater Correlation Coefficients and Standard error of measurement for Good Dog Carl (GDC) and Picnic

		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

All ICCs are positive and significant ($p < .05$)

CHAPTER 6

STUDY Ⅳ

Measurement of word retrieval in the discourse of persons with aphasia: Standard core lexicon item development and psychometric properties

INTRODUCTION

Discourse production is a complicated and multifaceted process, leading to scant use of discourse analysis in clinical settings. In recent years, there have been various opinions as to what the key elements are in successful, clinical outcome measures for discourse production. A general consensus was reached among aphasiologists. The first point relevant to clinical use of discourse analysis concerns a cost-benefit analysis (Coelho, 2007; Kintz & Wright, 2018; Wallace, Worrall, Rose, & Le Dorze, 2014). Clinicians generally provide assessment and treatment on the basis of a cost benefit analysis for their patients (Coelho, 2007). It is doubtful that in this situation, existing discourse measures are effective, compelling means of assessing discourse performance in a clinical setting. The second issue concerns discourse elicitation techniques (Coelho, 2007; de Riesthal & Diehl, 2018; Dietz & Boyle, 2018; Wallace, Worrall, Rose, & Le Dorze, 2018; Whitworth, 2018). Researchers have suggested considering which discourse elicitation techniques best represent communicative exchanges or best fit with specific outcome measures in the acknowledgement of a limited time in clinical settings. Coelho (2007) has noted that eliciting conversational discourse takes more than 15 minutes because of the involvement of a third party as a conversational partner. Another point concerns the time and training for the entire process of discourse analysis (Kintz & Wright, 2018; Kurland & Stokes, 2018). For example, discourse analysis requires collecting, transcribing, and analyzing language samples. A trained clinician generally requires more than four times the actual length of the discourse

sample just to complete the transcription process alone (Armstrong, Brady, Mackenzie, & Norrie, 2007; Boles & Bombard, 1998; Elia, Liles, Duffy, Coelho, & Belanger, 1994; Boles & Bombard, 1998). This timeframe excludes the time required for training and analysis, thus making many analyses impractical for use in clinical settings. The fourth issue concerns the reliability of the transcription and segmentation of discourse samples. The process of transcribing language samples orthographically has been reported to impede the achievement of high reliability of discourse analysis (McNeil, Doyle, Fossett, Park, & Goda, 2001; Olness, Gyger, & Thomas, 2012; Riesthal & Diehl, 2018). Moreover, it has been suggested that how discourse samples are segmented impacts outcomes, even with the same analysis (Kintz & Wright, 2018). A fifth point relevant to the clinical use of discourse analysis involves the lack of normative data to provide clinical guidance for assessment (de Riesthal & Diehl, 2018; Olness et al., 2012; Wallace et al., 2014). Normative data allow clinicians to interpret patients' performance, and further assess their linguistic changes following treatment.

A series of studies on the core lexicon measure have demonstrated that it can serve as an option to reconcile ecological validity with clinical usability in accordance with consensus discussed among researchers. As demonstrated in previous sections, the core lexicon measure is a means of assessment that is easily quantifiable, time-saving, and highly reliable, without preparatory work (e.g., transcription). Further, our comprehensive core lexicon measures including both content and function words have provided information as to how much and in what ways PWA' discourse performance differ from those of cognitively healthy adults. However, a methodological challenge for the development of a standard core lexicon list still exists. In the sections that follow, I provide methodological issues that should be addressed in order to pursue the development of a standard core lexicon set. First, I review the literature regarding the different discourse elicitation tasks that have been used in aphasia research. Then, I summarize the criteria used in previous studies to identify core lexicon items.

Discourse Elicitation Task

Many studies have focused on whether different elicitation tasks have a significant impact on the quality and quantity of language samples. Ulatowska, North, and Macaluso-Haynes (1981) investigated discourse performance on persons with aphasia (PWA) and neurologically intact adults (NIA) using two different discourse tasks (procedural and narratives) based on the idea that quality of language samples from narrative discourse is clearly different from that of procedural discourse. Ten PWA and 10 NIA participated in this study. For narrative discourse, recounts, sequential picture description, retelling of a specific story, and telling a summary and opinion of the story were included. For the procedural discourse, picture stimuli were provided to facilitate the nature of the task prior to participants stating the task. Participants were asked to describe how to do something in a series of steps, such as brushing their teeth, making sandwiches, and changing a tire. A comprehensive analysis of the language sample was conducted. Length of T-units, complexity of language, discourse grammar, and a rating system to evaluate the content and clarity were included as dependent variables. As a result, significant differences were found between PWA and NIA in terms of all dependent variables with the exception of number of T-unit for narrative discourse, and percentage of dependent clauses for procedural tasks. With respect to comparisons on discourse elicitation tasks, the control group produced more language for four dependent variables (words and clauses per T-unit, and percentage of dependent clauses and nonfinite clauses) out of six dependent variables on the narrative discourse tasks compared to the procedure discourse tasks. Similar results were found for the PWA, except for T-units. The PWA produced longer T-units for the procedural discourse tasks compared to the narrative discourse tasks.

In keeping with this idea, Hartley and Jensen, (1991) examined the effect of different discourse tasks (procedural and narrative) on productivity, content, and cohesion in persons

with closed-head injury (CHI) and NIA. The narrative tasks included a story retelling task and a story generation task based on a comic strip. The procedural task was to explain steps of how to buy groceries. The researchers analyzed discourse samples across 13 measures, and investigated the relationship between those discourse outcomes and cognitive and linguistic performance determined by the Western Aphasia Battery (WAB; Kertesz, 1982) and Wechsler Memory Scale (WMS; Wechsler, 1945). The NIA group consistently outperformed the CHI group on the discourse measures. Further, the CHI participants' performance as measured by the discourse measures did not differ across the two discourse tasks (story retelling and story generation); yet significant differences were found on seven of the 13 discourse measures for the NIA group. Several significant correlations were found among the discourse measures across the three tasks (story retelling, story generation, procedural discourse) and performance on the language and memory measures. The researchers concluded that discourse impairments varied in relation to type of discourse.

Olness, Ulatowska, Wertz, Thompson, and Auther (2002) examined the quantity and quality of discourse across different discourse elicitation tasks between African American and Caucasian individuals with and without aphasia. The picture stimuli consisted of two single pictures and one sequential picture scene. Analyses completed on the language samples included number of propositions, thematic content, ethnic dialect occurrence, and inclusion of feature characteristics of discourse type. Regardless of ethnic group, PWA produced fewer propositions when describing one of the single picture stimuli compared to cognitively healthy adults. Olness and colleagues reported qualitative differences between the single and sequential picture tasks. Both single pictures tasks elicited more descriptions of characters and actions compared to sequential picture tasks. Sequential picture tasks produced more temporal progression of events. The researchers interpreted the findings to suggest that

task selection is critical to consider when evaluating discourse abilities in both research and clinical use.

Coelho (2002) examined the effect of different story elicitation procedures (using multiple pictures and a single picture) on discourse abilities in CHI and NIA groups. For the multiple picture task, a pictured story consisting of 19 frames was provided. After sequentially viewing the pictures, participants were asked to tell the story without the stimuli. For the single picture task, a copy of a drawing describing an event (Norman Rockwell's painting, the Runaway) was provided to the participants. While looking at the drawing, participants had to explain what was happening. Dependent measures for both discourse elicitation tasks included utterance production (number of words per T-unit, number of subordinate clauses), cohesion (adequacy of cohesive marker), and story grammar (number of total episodes, proportion of t-units). The NIA group performed better compared to the CHI group for words per T-unit and T-units within the episodic structure. For both groups, words and subordinate clauses per T-unit were produced more frequently in the single picture task than in the multiple picture task, whereas greater cohesive adequacy and episodes were found in the multiple picture task than in the single picture task. Based on the findings, Coelho's concluded that the single picture task required participants to create the story content , thus resulting in lengthier utterances. Coelho also concluded that the restrictive plot structure in the multiple picture task facilitated participant's story grammar production compared to the single picture task.

Capilouto, Wright, and Wagovich (2005) examined the effect of discourse elicitation techniques on different discourse measures including correct information unit (CIU) and main event analyses in healthy younger and older adults. The stimuli consisted of four story-telling tasks from Nicholas and Brookshire (1993) - two single picture stimuli (Birthday Cake, and Cat in the Tree), and two sequential picture stimuli (Fight, and Directions). The

younger adults (YG) yielded more informative content than the older adults group (OG). For both groups, sequential pictures elicited greater proportion of main events conveyed compared to the single picture stimuli. The researchers suggested that the sequential pictures stimuli elicited discourse samples that included more temporal and causal information beyond simple descriptions of the characters and action, thus resulting in greater proportion of main events conveyed.

Wright and Capilouto (2009) used the same four discourse tasks as Capilouto and colleagues (2005) to compare linguistic performance of two groups of YG and OG on other discourse measures. These measures included information units (IU), lexical diversity (D), syntactic analysis, and main events. For the sequential picture task, percent of IUs produced and proportion of main events conveyed significantly correlated. For the single picture task, a significant, negative correlation was found between *D* and proportion of main events. Further, participants produced more lexically diverse discourse samples for the sequential picture stimuli compared to the single picture stimuli.

Fergadiotis, Wright, and Capilouto (2011) investigated the effect of discourse type on lexical diversity between two cognitively healthy groups (younger and older). Four different discourse tasks were used (procedures, eventcasts, story-telling, and recounts). For the procedural discourse, participants were instructed to explain the steps on how to make a peanut butter and jelly sandwich, and how to plant a flower in a garden. For the eventcasts, two single pictured scenes (The Birthday Cake and The Cat in the Tree) were provided, and participants were asked to tell a story with a beginning, middle, and end. For the story-telling task, a wordless picture book depicted a family of mice going on a picnic. Participants had to tell a story based on the pictures provided. For the recounts, three events (last weekend, last vacation, last Christmas) were suggested to elicit the language sample from the participants. The lexical diversity measure used was the *D* index (Malvern & Richards, 1997), which is an

estimate of individuals' range of vocabulary. D has been known to be robust to variations in length of language sample (McKee, Malvern, & Richards, 2000). D index was obtained using the voc-D program in CLAN (MacWhinney, 2000). Significant main effects were found for discourse type and age. For both groups, procedures resulted in the least lexically diverse samples; whereas, recounts had the greatest lexical diversity. Further, the OG produced greater lexical diversity for the procedures and recounts compared to YG.

Fergadiotis and Wright (2011) further investigated lexical diversity across different discourse elicitation tasks for adults with aphasia. Language samples from 25 PWA and 27 NIA were retrieved from AphasiaBank. Three types of discourse tasks (single pictures, sequential pictures, story-telling) included in the AphasiaBank protocol were selected. The single picture stimuli included Nicholas and Brookshire's "Cat Rescue" (Nicholas & Brookshire, 1993) and a photograph by Anni Wells (Rubin & Newton, 2003). The sequential picture stimuli included two cartoon strips of four-frames and six-frames (Broken Window, Umbrella). Language samples of the Cinderella storytelling were used. Significant interactions between groups and discourse tasks were found. For both groups, single picture descriptions produced the least lexically diverse samples. Significant group differences were found for all three discourse tasks with the NIA group producing greater, lexically diverse stories compared to the PWA group. Groups differed for the discourse elicitation task that resulted in the greatest lexical diversity – for the NIA group it was the story telling task and for the PWA group it was the sequential pictures stimuli. The researchers concluded that PWA's high lexical diversity in the sequential pictures was attributed to inherent characteristics of the stimuli that provide abundant story elements. Further, opposed to the single picture task, multiple, sequential pictures that provide story structure serve as a cognitive map, prompting PWA to produce more lexical items. This account is consistent

with previous studies where sequential pictures elicit a better quality of discourse from adults with acquired language disorders (Coelho, 2002; Olness et al., 2002).

Results from these studies demonstrate how critically important selection of discourse elicitation task is. Further, different cognitive and linguistic demands are imposed on speakers depending on the type of task (Bliss & McCabe, 2006; Brady, Armstrong, & Mackenzie, 2005; Nicholas & Brookshire, 1993). In the field of aphasiology, various discourse elicitation tasks have been used, such as procedures, eventcasts, recounts, and storytelling. For example, researchers have asked participants to describe step-by-step descriptions of an activity (e.g., how to make a peanut butter and jelly sandwich) (e.g., Bartels-Tobin & Hinckley, 2005; Rider, Wright, Marshall, & Page, 2008). Some researchers used eventcasts, in which participants describe a scene (e.g., Cat in the Tree) (Capilouto et al., 2005; Fergadiotis et al., 2011; Fergadiotis & Wright, 2011; Olness et al., 2002; Wright & Capilouto, 2009). Other researchers asked participants to tell their personal events that happened in the past (e.g., what one did last Christmas) (Armstrong et al., 2007; Fergadiotis et al., 2011; Wright & Capilouto, 2009) and/or to tell a familiar story (e.g., Cinderella) (e.g., Coelho, 2002; MacWhinney et al., 2011; Webster, Franklin, & Howard, 2007).

There is some disagreement about which type of discourse is more demanding, and more appropriate for a clinical decision. In the literature, procedural discourse (e.g., how to make a peanut butter and jelly sandwich) is a widely used task because of its simplicity and reduced individual variability (Ulatowska, North, & Macaluso-Haynes, 1981). Alternatively, procedural discourse tasks used typically include known steps and sequences that have been overlearned and are not cognitively demanding; thus limiting ability to capture disruptions in discourse (Weiss, 2012).

Recounts is preferable to researchers who are interested in spontaneous, natural speech that involved in speakers' emotional reaction (Law, Kong, & Lai, 2018). It was also

reported discourse appears to elicit the most diverse vocabulary from PWA and cognitively healthy individuals (Fergadiotis & Wright, 2011); however, it is associated with the ability to activate long-term memory to retrieve details. With use of recounts, though this elicitation task may produce diverse language samples, it is often burdensome for clinicians because of the inability to evaluate accuracy of information conveyed.

Compared to procedural and recounts, narrative discourses obtained from wordless picture books or a scene of activities (eventcasts) is more appropriate to obtain lexically diverse language samples (Fergadiotis, 2011; Fergadiotis & Wright, 2011). During the task, pictorial support is provided, which evokes concrete and high-imageability words (Grosjean, 1980), as well as, visual imagery of the actions (Fergadiotis, 2011). A critical difference between wordless picture books and eventcasts is level of complexity. Fergadiotis (2011) reported that eventcasts have simpler story grammar with fewer episodes. In contrast, using multiple picture stimuli, such as wordless pictures books, to elicit language samples typically provides story elements such as setting, characters, problems, and actions. Major events occur in a specific time, place, and social environment. As the story proceeds, main characters arrive at the highest peak of tension, provoking an emotional response. Following the story structure, speakers need to select lexical items to describe characters and events.

Telling a familiar story, such as the Cinderella story, is another way to elicit narrative discourse. Although telling a familiar story may elicit abundant language samples, cognitive confounds are more likely to occur compared to telling a new story. When producing a story that individuals already know, speakers' cognitive resources are likely to be divided to retrieve stored representation from memory. It would not be an optimal type of story that actually taps narrative skills. The second concern is that past experiences with stories can also affect discourse performance (Gazella & Stockman, 2003). For example, the Cinderella story is culturally entrenched, and transmits cultural history and values. Schematic organization of

the story can vary across individuals with or without a multicultural background. Whether or not speakers experienced formal schooling can affect the quality and length of language samples. Marshall and Cairns (2005) suggested that during story retelling, individuals select events and details to describe the story effectively as they recall. This also has been supported by Carragher, Sage, and Conroy (2015). When extraneous aspects are better controlled, more reliable language samples can be acquired. Thus, instead of using the familiar story, a novel story that participants are not familiar with should be standard practice.

Criteria for Core Lexicon Items

A comprehensive review of the core lexicon analysis literature is provided in Chapter 2. Summarized in the following section how previous core lexicon analysis research identified lexical items. Core lexicon measures are a relatively new method and, as such, few works have directly applied the analysis to discourse produced by adults with aphasia. Not surprisingly, studies focusing on developing a core lexicon measure have used different discourse elicitation techniques and criteria to determine lexical items. In order to take a step towards extending the current methodology to develop a standard core lexicon set for discourse evaluation, the two issues are reviewed.

Core lexicon analysis is a lexicon-based analysis that provides a checklist consisting of critical lexical items required to deliver semantically meaningful and coherent discourse. In this sense, it is clear that the construction of core lexicon lists would be directly related to specificity and sensitivity of the measure. In regards to the criteria to select core lexicon items from normative data, converging evidence among previous studies is limited. For example, in previous studies, inclusion criterion required at least 50% of the core lexical items to be produced by the control participants (Dalton & Richardson, 2015; Dillow, 2013). However, Fromm and colleagues (2013) did not specify an inclusionary criterion and extracted the top 10 core lexicon items for verbs and nouns. MacWhinney, Fromm and colleagues (2010) also

did not stipulate a criterion and extracted the top 10 verbs and nouns (approximately 20% of the entire lexicon items). Moreover, only Dalton and Richardson explained the rationale for their inclusionary criterion. Their criterion of lexical items produced by greater than 50% of the sampling cohort was established based on Roger Brown's language stages of typical language development in children (Brown, 1973). Despite the fact that Brown's stages are extensively used as a measure of determining whether language is delayed in children, it is not ideal that a principle stemming from a standard pattern for children is applicable to research in acquired language disorders. Additionally, Brown (1973) analyzed data from three children for his longitudinal study, and one child dropped out of the study after a year.

As demonstrated in Studies 1 to 3, our core lexicon lists were developed based on language samples elicited by wordless picture books, and consisted of the most frequently occurring 25 items by word class. Previously, the field of computational linguistics has been interested in corpus studies generally based on writing. In much research involving a corpus linguistics approach, frequently occurring words were extracted and used to examine word importance in a given text (Baker, 2004; Finch & Chater, 1992; Gauch, Wang, & Rachakonda, 1999; Pustejovsky, Anick, & Bergler, 1993). While there are such precedents for defining text from a frequency list, the cut-off was arbitrarily selected, with ease of use being the most important factor in that decision (Gottron, 2009). In sum, the lack of statistical guidance for clinically manageable criteria may pose a serious challenge to the development of a standard core lexicon list and the potential use of this measure in clinical settings.

The purpose of the current study, then, is to: 1) determine a better criterion for core lexicon list development; and 2) examine psychometric properties of the lexical items in a standard set from a large database of PWA. The specific aims of the study are to (1) examine whether a specific approach (i.e., frequency or percentage) is more valid to identify core lexicon items; and (2) examine if it is possible to employ a universal core lexicon list for

clinical purposes. The results of the current study are unpredictable since core lexicon studies and development of core lexicon lists are in a nascent stage. However, for Study Aim 1 (SA1), it is reasonable to assume that there is minimal difference between the core lexicon lists created by the two different criteria (frequency vs percentage) for some word classes such as core nouns, verbs, and function words, given that we have somewhat sufficient language samples which are the methodological basis of core lexicon lists. In contrast, it is predicted that statistically meaningful measurement invariances across the two core lexicon lists will be found for adjectives and adverbs because the number of modifiers produced by speakers are scant. For Study Aim 2 (SA2), clinical feasibility and application will be evaluated by using an item response theory (IRT) approach in order to develop a theoretically precise, reliable, and valid core lexicon list. Through these analyses, I expect to have a finalized, short form list that is specific for quantifying word retrieval ability at the discourse level in any context. To the best of our knowledge, this is the first attempt to demonstrate the development and validation process of the core lexicon measure's psychometric properties. This will help achieve more effective outcomes with greater measurement precision.

METHOD

Study Population

For SA1, language samples from 470 cognitively healthy adults (Wright & Capilouto, 2017) were used. For SA2, a large database of PWA (N =272) from the AphasiaBank (MacWhinney, 2000) was used (Table 6-1).

Statistical Approach

For SA1, two different core lexicon lists were created based on the different approaches to defining lexical items. These included core lexicon frequency (CLF) and core lexicon percent (CLP). Then, language samples from 470 cognitively healthy adults were

used to compare CLF and CLP for core lexicon measures. Confirmatory factor analysis (CFA) was computed because it is possible to estimate parameters to draw information regarding the measurement invariance. Compelled by the result from SA1, a better criterion among two approaches was applied to SA2 analysis. Item response theory (IRT) was computed to investigate psychometric properties of these lexical items in the core lexicon lists.

Developing a Standard Core Lexicon Set

A paradigm shift is occurring in aphasia research – researchers are moving away from chaotic discourse measures to standardized outcome measures for clinical and research settings (Dietz & Boyle, 2018b). Thus, it is imperative to take a step towards extending current methodology to address issues of ecological validity and clinical feasibility. Since the core lexicon has proven to be reliable in previous research (Dalton & Richardson, 2015; Dillow, 2013; Fromm et al., 2013; MacWhinney et al., 2010), and our core lexicon measure has been developed using a large number of controls, it is appropriate for it to be the first step in developing a standard outcome for discourse evaluation. However, some issues need to be considered, such as discourse elicitation technique and word class, which have been addressed in previous sections. Another note of concern in developing a standard core lexicon set is whether or not the standard core lexicon set is well developed based on grounded theoretical and practical motivations. Thus, in the following section, theoretical rationale and empirical evidence will be provided for refinement of our core lexicon measure and further development of a standard core lexicon set by word class.

Nouns. Admittedly, a standard core lexicon list for nouns does not seem to be plausible because target nouns are tightly linked to objects depicted in picture stimuli. For example, MacWhinney and colleagues (2010) shared the top 10 nouns and verbs extracted from cognitively healthy adults' telling of the Cinderella story. The most frequently occurring

nouns were associated with story-specific words such as Cinderella, prince, stepmother, stepsister, slipper, and fairy. Fromm and colleagues (2013) reported the top 10 nouns for a procedural task in which participants explained how to make peanut butter and jelly sandwich. The controls also produced task-specific words such as bread, butter, peanut, jelly, slice, and knife. As such, to achieve the goal of a standard core noun set, collapsed data of language samples retrieved from different narrative tasks are inevitable.

Verbs. Numerous studies have focused on verb production in different contexts (e.g., single word, sentence, and discourse level). These investigations originate from multiple theories that verb usage is affected by lexical and grammatical aspects in language processing. For example, some researchers have demonstrated that PWA's difficulty in retrieving verbs has been attributed to imageability and phonemic length of verbs (e.g., Bird, Howard, & Franklin, 2003; Mätzig, Druks, Masterson, & Vigliocco, 2009; Shapiro & Levine, 1990). Others have examined verb argument structure as another factor affecting verb retrieval (e.g., Thompson, Lange, Schneider, & Shapiro, 1997). Another line of studies has investigated the effects of semantic weights on the retrieval of verbs (Berndt, Haendiges, Mitchum, & Sandson, 1997; Gordon, 2008). It is difficult to determine which of these attributes is theoretically associated with the development of a core verb set. Possibly, the differential impairments of light and heavy verbs produced in PWA support the potential of clinical feasibility of core lexicon measures.

Jesperson (1965) initially coined the terms "light verb" and "heavy verb" in his analysis of English sentence construction of verb and noun phrases. By definition, light verbs are semantically unspecified; they present general meaning that can be widely used in various contexts. Examples of light verbs are: *do*, *go*, *get*, *make*, and *take*. Heavy verbs are defined as semantically complex verbs that include specific meaning; for example, *kick*, *drink*, *fly*, and *read.* These verbs carry a concrete meaning used in a smaller range of events. More

specifically, Jespersen (1965) suggested that light verbs play a role in marking person and tense in sentences, instead of carrying major semantic burden. Semantic burden is generally carried by the nouns. For example, two sentences "John looked at the boy", and "John took a look at the boy" roughly deliver an equivalent message (Miyamoto, 2000). In this context, *take* is considered a light verb, but *look* is not included in the light verb category (Clark, 1978), suggesting that light verbs are used to express general purpose along with abstract particles (e.g., up, away, and off) in the majority of cases (Maouene, Laakso, & Smith, 2011). Pinker (2013) suggested light verbs are translational cases and lie somewhere between open and closed class words.

Similar to the linguists' definition of light verbs, the inclusion and/or exclusion of some verbs have differed across aphasia studies. An earlier explorative study of verb production in aphasia narratives classified verbs into light and heavy verbs based on the frequency of language samples (Berndt et al., 1997). High frequency verbs that are considered as light verbs include *come, go, make, take, get, give, do, have, be.* In Berndt and colleagues' (1997) study, a pattern of verb production was found. Six of seven participants with non-agrammatic aphasia produced a higher proportion of heavy verbs, and all participants with agrammatic aphasia produced a higher proportion of light verbs in their utterance and discourse production tasks.

Later studies have reported a dissociation between light and heavy verb usage in connected speech (e.g., Breedin, Saffran, & Schwartz, 1998; Kim & Thompson, 2004). Breedin, Saffran, and Schwartz (1998) examined verb production in aphasia narratives, using light and heavy verbs paired for story completion tasks. They found that study participants with agrammatic aphasia produced more heavy verbs than light verbs (i.e., *bring, come, get, give, go, have, make, move, put, and take*), whereas the participants with non-agrammatic aphasia show the opposite pattern. In Kim and Thompson's (2004) study, light verbs (i.e., *be,*

come, do, get, give, go, have, make, move, put, and take) were determined based on the discussion in three previous studies (Breedin et al., 1989; Jespersen, 1965; Pinker, 1989). Adopting the same light verb lists as Kim and Thompson, Gordon (2008) reported that the proportion of light and heavy verb usage discriminated persons with fluent aphasia from persons with non-fluent aphasia.

The initial study of core lexicon analysis (MacWhinney et al., 2010) compared the top 10 verbs extracted from their study to light verbs that Gordon (2008) tracked. They reported that seven of their core verbs were in common with light verbs identified by Gordon (2008). As another example, when reviewing the GDC list for the 20s age group, eight of Gordon's (2008) light verbs occurred in Kim and colleagues' (2019) study. Copula was not included. For the Picnic story, *give, take, make, move,* and *be* were not included in the 20's age group list. The other verbs included in the core lexicon lists consisted of comparatively heavy verbs from the point of semantic weights. As noted above, core verbs produced cannot be completely interpreted within a framework of light verb usage in aphasia narratives. However, the procedures to create core lexicon lists seem analogous to the determination of light verbs based on the frequency (Berndt et al., 1997).

Modifiers. Modifiers have not been consistently considered in discourse-level word retrieval; thus it may be that the approach of creating a standard core lexicon list for modifiers does not ensure usability and feasibility of the list. However, knowing how much the person with aphasia's lexical usage is deviated from normal patterns may provide broader knowledge on preserved and impaired patients' ability for an effective intervention, and further potential avenues of specified clinical treatment.

In a language acquisition study, Tingley, Gleason, and Hooshyar (1994) noted that words referring to perception (soft, dark), and physiological states (hungry, sleepy) are relatively easy to learn. Given that some factors, such as age of acquisition and imageability,

affect the production of words in other word classes, it is reasonable to assume that modifiers (especially adjectives) related to perception and physiological states may be easier to produce and may be eligible to be core adjectives as a standard measure.

More recently, motivated by a previous study (i.e., Menn, Obler, & Miceli, 1990) Meltzer-Asscher and Thompson (2014) analyzed narratives of the Cinderella story elicited from 14 participants with agrammatic aphasia and 14 control participants. Meltzer-Asscher and Thompson (2014) primarily focused on adjective production rate, production of predicative and attributive adjectives, and production of adjectives with complex argument structure. Additionally, they approached adjective use in narratives within a framework in which word class production is affected by multiple dimensions, such as imageability, age of acquisition, and arguments structure. Participants were asked to rate the imageability of adjectives produced by both groups and compared the imageability of the adjectives in the narratives of persons with agrammatic aphasia and controls. Meltzer-Asscher and Thompson did not find significant differences between the two groups, and persons with agrammatic aphasia produced a greater proportion of predicate adjectives and a fewer proportion of attributive adjectives in their narratives compared to the control group. Although these findings are hard to draw strong conclusions from due to the small sample size, the researchers attempted to explain their findings within a framework of *output economy strategy* for agrammatic speakers. Since speech is effortful for persons with agrammatic aphasia, they tend to avoid producing syntactically complex sentences that contain attributive adjectives. The overproduction of predicative adjectives was considered to be reflective of verb deficits that agrammatic speakers generally experience. Findings that persons with agrammatic aphasia produced adjectives in narratives similar to that of control participants seems unlikely and contradicts findings from a previous investigation (Varley & Siegal, 2000).

In another study, Sarno, Postman, Cho, and Norman (2005) conducted a longitudinal study and examined language gains for verbs, nouns, adjectives, and adverbs in participants with fluent (N = 11) and non-fluent aphasia (N = 7) following language therapy. A comprehensive treatment program consisting of individual and group interventions was administered for six weeks. The interventions were devised to enhance linguistic deficits, pragmatic skills, and functional communication in learning strategies. During the course of treatment, the percentage of modifiers (adjectives and adverbs) increased, whereas the percentage of nouns and verbs decreased. They concluded that increased use of adjectives and adverbs resulted in improved lexical diversity for PWA. The authors did not statistically compare results between the fluent and non-fluent aphasia groups.

Within the context of this study, a question arises as to what adjectives constitute a core set of adjectives. To the best of our knowledge, this is the first study to develop a measure quantifying use of modifiers in discourse. Currently, the relationship between modifier usage and aphasia subtypes or aphasia severity is still nebulous. As an explorative approach, alleviating the task effects with collapsed discourse samples of two discourse tasks (GDC and Picnic) will assist in addressing the specific aims.

Function Words. Development of a standard core function word set seems to be more reasonable compared to other word classes. English speakers produce a limited number of function words (Chung & Pennebaker, 2011). Chung and Pennebaker (2011) collected 95,000 text samples from 80,000 different people using various themes such as descriptions of an object, event, daily routine, or personal accounts for emotional events. They identified the top 20 most commonly used function words from their text archive. The top 10 words among the 20 function words are also the most frequently used words and account for 20% of the words spoken. Moreover, in our previous study (Kim et al., 2016, 2017), function word usage was not heterogeneous across the age cohorts (20s, 30s, 40s, 50s, 60s, 70s, and 80s).

Contrary to content word usage, the agreement range was comparatively tight for function words (range: 84% - 100%) across the age spectrum compared to percent agreement for content words (verb range: 64% - 92%, noun range: 56% - 92%, adjective range: 60% - 92%, adverb range: 72% - 92%) (Kim et al., 2016, 2017). However, it is possible that some function word usage, such as pronouns, are task dependent. In the wordless picture books used in Kim and colleagues' studies, no character included would be appropriate for eliciting the pronoun "*I*". In Chung and Pennebaker's (2011) list, the pronouns "*I"* and "*My"* were included because of how the language samples were collected. Therefore, characters in picture stimuli influence the presence of specific pronouns in language samples and should be acknowledged when developing a standard list.

In summary, an initial step to developing a standard set of core lexicon is to mitigate story-specific words. To do this, collapsed data across two different narrative tasks will be utilized. This will enable us to extract core lexicon items in a decontextualized manner.

Evaluating the quality of measurement instruments

Issues related to the quality of measurement instruments are receiving increasing attention in research. In a traditional concept of scientific measurement, measurement is defined as the estimation of a quantitative attribute of abstract concepts (Michell, 1997). In health care and social science research, measurement involves the operationalization of theoretical constructs and the development and application of instruments to quantify specified phenomena (Kimberlin & Winterstein, 2008). A great deal of research in the field of speech-language pathology involves quantifying the behaviors of patients that cannot be directly measured. For example, correct information units (CIU; Nicholas & Brookshire, 1993) are operationally defined as an information unit that is relevant to the stimuli in connected speech. Previously, researchers had tried to prove whether CIU actually measures what it purports to measure and were successful in doing so (e.g., Brookshire & Nicholas,

1994; Doyle, Goda, & Spencer, 1995; Oelschlaeger & Thorne, 1999). This allowed CIU to be used as one of acceptable outcome measure for informativeness in clinical practice and research. In other words, a central issue in constructing a new language outcome measure or test is whether the intended linguistic functions will be appropriately measured by them. Accordingly, reliability and validity are of particular relevance to aphasia language tests because these measures reflect not only soundness of the test, but the test's ability to discriminate severity.

Three types of reliability that validate the quality of the test have been suggested by Ivanova and Hallowell (2013): Internal consistency, test-retest stability, and inter-rater reliability. Internal consistency represents the constancy of results across items, which is generally estimated using Cronbach's alpha. Test-retest reliability reflects the stability of results across time. Inter-rater reliability is a measure of consistency between different examiners administering a test, which is estimated by intraclass correlation (ICC). For dichotomous items, Kuder Richardson (KR-20) is more appropriate to use instead of Cronbach's alpha.

Validity is equally important for creation of a new language test. Kimberlin and Winterstein (2008) pointed to the importance of validity, stating that an instrument that is reliable is not always found to be valid. The term validity refers to the degree to which the intended linguistic functions are measured and the specific inferences about participant groups are appropriate. Validity has been investigated from four different perspectives relevant to aphasia language batteries (Ivanova & Hollowell, 2013): Face validity; content validity; concurrent validity; and construct validity. Face validity refers to the degree to which tests appear to measure what it was designed for. It is a subjective judgement by test administrators. Content validity provides evidence about how well a test measures the domain of functions intended to be measured. Concurrent validity represents the relationship

between the score on a test and scores on existing tests that are theoretically considered to index the same underlying behaviors. Lastly, construct validity pertains to the extent to which tests actually measure what they were intended to measure. In the test development process, validity is a requirement, as different aspects of validity provide additional, strong evidence of the interpretation of test scores. The cornerstone of developing a standard set of core lexicon at this point rests in establishing construct validity of theoretical constructs. In the following section,

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is a type of structural equation modeling (SEM) often used to test the dimensionality of tests. CFA approaches provide a strong, analytic framework to account for relationships between abstract concepts (unobservable variables) and observed scores (Brown, 2014). A construct is a theoretical concept, and typically associated with multiple measures. For example, word retrieval deficits are multidimensional constructs defined by various forms of language difficulty (e.g., impairments of lexical semantic processing). Unlike physical quantities such as height and weight (Borsboom, Mellenbergh, & van Heerden, 2004), word retrieval deficits cannot be directly observed or measured, but can be estimated by using outcome measures. Conceptually, core lexicon analysis is devised to capture speakers' word retrieval ability at the discourse level using pre-determined lexical items. That is, word retrieval ability can be conceptualized as a latent construct within the SEM framework.

In studies that have focused on core lexicon measures, two different criteria to identify core lexical items have mainly been used (Dalton & Richardson, 2015; Kim et al., 2019). It is unknown which criteria would be better to construct core lexicon lists for more accurately measuring the underlying word retrieval ability in PWA's narrative discourse. As such, core lexicon sets constructed by two different criteria are conceptualized as the

measurement tools to obtain an observable score: one is a core lexicon set based on frequency (CLF), and the other is a core lexicon set based on percent usage (CLP).

In the context of this study, longitudinal measurement invariance is appropriate to use in that two factors (CLF and CLP) represent a similar construct for the same group. A path diagram in Figure 6.1 shows how each construct is measured by five observed indicators (e.g., nouns, verbs, adjectives, adverbs, function words). For each measure (i.e. CLF, CLP), higher scores reflect a greater ability to retrieve typical words at the discourse level. An initial step of the analysis is to determine whether among the two approaches, the latent constructs of word retrieval ability have the same definition. Based on the path diagram, there are a total of 65 elements - 55 variances and covariances and 10 means. Ten observed means and 12 unknown parameters (10 indicator intercepts, 2 latent factor means) exist in the CFA structure. Considering that the model is under-identified, we will fix the latent mean to zero for the constructs to identify the mean structure. To decide which approach is better to use, measurement parameters were compared across models to examine measurement invariances. Testing for measurement invariance involves comparing models in which measurement parameters (e.g. factor loadings) are systematically held invariant across measures. The first model (configural model) allows all measurement parameters to be freely estimated. Subsequent models are nested under the previous models with restrictions. For example, a strong invariance model to test intercepts across measures entails a model in which factor loadings are restricted to be equal across measures. Ultimately, one approach which reveals higher factor loadings and lower measurement errors will be selected as a better criterion. At each step, model fit indices will be used: Chi-Square goodness of fit test; the Root Mean Square Error of Approximation (RMSEA; Steiger, 1990); the Comparative Fit Index (CFI; Bentler, 1990); and Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), following the cut off values suggested by Hu and Bentler (1999).

Item Response Theory

For test development, it is also important to investigate characteristics of individual items in the test instrument. In the current study, we implement a modern measurement theory approach, item response theory (IRT), which is distinct from classical test theory (CTT). Although the concept of IRT models is not new in the field of speech-language pathology, we are more familiar with CTT. In CTT, a level of an attribute is estimated as the sum (Nunally & Bernstein, 1994). It is assumed that every person has a true score on an item, which consists of true score and some measurement error. In this concept, all items of an instrument should be administered to obtain a valid score (Baylor et al., 2011). Accordingly, more items in a test instrument are logically associated with higher reliability of the test score. Test scores need to be interpreted by comparing patients to their respective counterparts with a sum of scores.

Advantages of IRT models over CTT are apparent. IRT allows adaptive testing, in which the difficulty level of test items can be tailored for each individual. This leads to greater measurement efficiency by reducing response burden and time for test administration (Cook, O'Malley, & Roddey, 2005). For the interpretation of test scores, it is possible to understand an individual's ability through the use of the logit scale (the scale of measurement in IRT), rather than depending on the particular person used for comparisons.

IRT models were first introduced in the 1950s and 60s by Fred Lord and Alan Birnbaum (Birnbaum, 1968; Bock. 1997; Lord & Novick. 1968) and increasingly applied in psychology and health-related fields over the past 15 years. George Rasch also developed a similar model which is mathematically identical to the simplest IRT model, the one-parameter logistic (1-PL) model. There are also more complex models, the two-parameter logistic (2-PL) model and three-parameter logistic (3-PL) model. The 1-PL model assumes that all items are equally difficult, and that individuals who are low in the attribute have little

chance of guessing the correct answer. The 2-PL model estimates both item difficulty and discrimination. The 3-PL model includes not only item difficulty and item discrimination, but also a guessing parameter, considering that some respondents may have a higher propensity to guess correct answers for items. In the current study, I implemented the 2-PL model to design a final standard set of the core lexicon measure. The 2-PL model was computed using statistical packages (e.g., Itm, Rizopoulous, 2006) designed for IRT analysis within R (R Core Team 2013).

IRT models entail two assumptions: unidimensionality and local independence. In IRT models, all items are characterized by a single dominant underlying dimension (unidimensionality). The local independence assumption requires that item response on any given item is not correlated with item response on any other item. Once the assumptions are met, item calibrations are estimated to create a final standard set of core lexicon measures. Then, items can be deleted based on the item difficulty hierarchy. The final model will be selected in accordance with the Akaike's Information Criterion (AIC: Akaike, 1974) and Schwarz's Bayesian Information Criterion (BIC: Schwarz, 1978) statistics. Figure 6.2 presents a flowchart of the procedures.

RESULTS

Measurement Quality of Core Lexicon

To test which of the two methods is better to use as a criterion for core lexicon measures, word retrieval ability was conceptualized as a latent variable, and its relationship with five observed variables (nouns, verbs, adjectives, adverbs, and function words) was modeled across two different approaches using CFA. Measurement invariance and structural invariance were examined by story (GDC and Picnic) using the *lavaan* package (Rosseel, 2012) in R (R Core Team, 2019). Prior to the CFA analysis, it was confirmed that no missing

data exist in data set. A configural invariance model was initially specified to test whether or not the same items measure our construct across two approaches by story. All factor means were fixed to 0 and all factor variances were fixed to 1 for identification and scale setting. Figure 6.1 depicts the initial proposed model. For GDC, the configural model fit was acceptable. Thus, the analysis proceeded by constraining the factor loadings to be equal across two approaches. To test if the factor loadings are equivalent across two approaches, the weak invariance model builds upon the configural model. The factor variance was fixed to 1 at CFA but was freely estimated at CLP. The weak invariance model had acceptable fit with mediocre to acceptable fit for absolute fit indices and acceptable to close fit for the relative fit indices. Chi-square differences were found between the configural model and weak invariance model with a large change in CFI, $\Delta\chi^2(4) = 36.751$, $p < .001$, $\Delta CFI = .011$, indicating that all factor loadings were not equivalent across two approaches. To assess the extent of measurement non-invariance, each factor loading was tested individually by comparing a model with the factor loading freely estimated to a model with the factor loading constrained to be equal across approaches. Results indicated non-invariance in two out of five factor loadings: CLF- adjectives, & CLP-adjectives, CLF-function words & CLP-function words. After freeing the factor loadings, the model fit became better and the weak invariance held. Then, to determine if the intercepts were invariant across two approaches, strong invariance model was tested. The strong invariance model fit significantly worse than the modified weak invariance model. The results from the nested model comparisons are provided in Table 6.2. The modification indices were examined to improve model fit. Results indicated non-invariance in four out of five item intercepts except adverbs. After doing so, the strong invariance model was significant, χ^2 (32, N=470) $= 993.998$, $p <$ .001. However, the model fit was poor for all model fits. The results indicated that item

intercepts were not invariant across both approaches. CLP had larger intercepts compared to CLF. Since factor loadings are invariant and the results did not provide a definitive evidence to choose one method over the other, residual invariance models were tested. The residual invariance model fit significantly worse than the final weak invariance model. The modification indices suggested except function word-related residuals, all residuals were not invariant across two approaches. CLP had larger residuals compared to CLF. After freeing four out of five residuals, the partial residual invariance model fit significantly better than the previous residual invariance model and did not fit worse than the final weak invariance model, $\Delta\chi^2(1) = .062562$, $p = .8025$, $\Delta CFI = .0003265$. Table 6.2 summarizes the fit indices for configural, weak, strong, and residual invariance models. The standardized parameters of the model can be seen in Figure 6.3.

For Picnic, the configural model fit was acceptable. The model fit was close for absolute and relative fit indices. Thus, the analysis proceeded by constraining the factor loadings to be equal across two approaches. To test if the factor loadings are equivalent across two approaches, the weak invariance model builds upon the configural model. The factor variance was fixed to 1 at CFA but was freely estimated at CLP. The weak invariance model was significant with acceptable fit for and close fit for relative fit indices. Chi-square differences were found between the configural model and weak invariance model with a large change in CFI, $\Delta\chi^2(4) = 48.421$, $p < .001$, $\Delta CFI = .0195$, indicating that all factor loadings were not equivalent across two approaches. To assess the extent of measurement non-invariance, each factor loading was tested individually by comparing a model with the factor loading freely estimated to a model with the factor loading constrained to be equal across approaches. Results indicated non-invariance in two out of five factor loadings: CLF-function words & CLP-function words, CLF-nouns & CLP-nouns. After sequentially freeing the

factor loadings, the model fit became better and the weak invariance held. Then, to determine if the intercepts were invariant across two approaches, strong invariance model was tested. The strong invariance model fit significantly worse than the modified weak invariance model. The results from the nested model comparisons are provided in Table 6.3. The modification indices were examined to improve model fit. Results indicated non-invariance in four out of five item intercepts except verbs. After doing so, the strong invariance model was significant, χ^2 (31, N=470) $= 47.521,\ p <$ .05. Model fits are close for all of model indices. Overall, since some factor loadings are invariant and the results did not provide a definitive evidence to choose one method over the other, residual invariance models were tested. The residual invariance model fit significantly worse than the final weak invariance model. The modification indices suggested except verb word-related residuals, all residuals were invariant across two approaches. CLP had larger residuals compared to CLF. After freeing four out of five residuals, the partial residual invariance model fit significantly better than the previous residual invariance model with close fit for absolute and relative fit indices. The model did not fit worse than the final weak invariance model, $\Delta\chi^2(1) = 2.8801$, $p = .0897$, $\Delta CFI = .0008$. Table 6.3 summarizes the fit indices for configural, weak, strong, and residual invariance models. The standardized parameters of the model can be seen in Figure 6.4.

Additionally, the variance invariance model was computed to see if variances differed across two approaches by story. $\Delta\chi^2$ was significant for weak invariance and latent variance invariance model, $\Delta\chi^2(1) = 5.0962$, $p <.01$, indicating that variances differed across the two approaches for GDC. Variability in the scores is larger in the percentage criterion compared to the frequency criterion. The model fit was close fit for SRMR and relative fit indices, and acceptable fit for RMSEA. For Picnic, $\Delta\chi^2$ was not significant for the weak

invariance and latent variance invariance model, $\Delta\chi^2(1) = 0.1056$, $p = 0.7452$, indicating that variances do not differ across the two approaches for Picnic.

IRT Model Assessment

Based on the results of SA1 that frequency is the better criterion to identify lexical items that construct core lexicon measures, core lexicon items for the IRT analyses were determined using the frequency criterion. Prior to item calibration, data were inspected to determine that a minimum of 1% of responses fell within each of the response categories or a maximum of 90% of missing data. Through this process, few items were deleted for nouns, verbs, adjectives, and adverbs. Then, the assumption of local independence using Pearson's chi-square statistics (Hanlbleton, Swaminathan, & Rogers, 1991) was tested. In traditional usage, any items that have a chi-square value of more than 3.5 do not meet the presumption of local independence with 95% confidence. As such, items that are locally dependent were removed from the data. For the second assumption of unidimensionality, it is assumed that only a single construct is being measured by core lexicon checklists because core lexicon lists for five word classes were established using the same narrative samples.

IRT model-fit assessment showed that the data had a significantly better fit to the 2-PL model than to the 1-PL model for verbs, nouns, adverbs, and function words, but not for adjectives (See Table 6.4). With respect to the decision of model selection, models that present misfit can be usable and sometimes can be preferred in IRT analysis for practical reasons (Zhao & Hambleton, 2017). Given the consistency of analysis among different word classes and the exploratory nature of the study, the 2-PL model was computed for all five word classes.

The 2-PL model was computed to calibrate item difficulty and discrimination of lexical items included in core lexicon measures. For the discrimination parameter, higher values are associated with items that are better able to discriminate those who perform well

and those who does not perform well. Thus, more discriminating items provide greater information about examinees' responses compared to less discriminating items. For the difficulty parameter, higher values indicate a higher proportion of examinees who provide a correct answer. Thus, higher values are associated with items that are easier to examinees. Most difficulty estimates fall between -2 and 2. Following the selection criteria, final core lexicon items were determined. Further, considering that different discourse stimuli were used for developing core lexicon lists, other factors such as context and missing data for items were considered for the final decision. More specific information is provided in the Figure 6.5.

As a result, consistent with expectations, function words demonstrated a good distribution of item difficulty and discrimination. Item discrimination parameters ranged from 1.11 (His) to 6.02 (And). These large item discrimination parameters indicate the core function word items discriminate examinees' traits very well (Baker, 2001). Item difficulty parameters ranged from -1.52 (And) to 1.53 (His). For verbs and adverbs, the possibility of potential development of universal lists was found, despite the limited range of item difficulty. Specifically, item discrimination and difficulty parameters for verbs ranged from 0.51 (Know) to 3.74 (Have), and from -0.84 (Do) to 1.56 (Leave), respectively. Item discrimination and difficulty parameters for adverbs ranged from 0.81 (Very) to 2.74 (Even), and from 1.04 (Back) to 1.92 (Where), respectively. However, nouns and adjectives did not fit into the model. Results with item difficulty and discrimination are provided in Appendix 6.A. The final, universal core lexicon items consisting of 9 verbs, 7 adverbs, and 10 function words are listed in the Appendix 6.B.

DISCUSSION

The purpose of this study was to contribute information about the measurement of core

lexicon measures, and about potential development and use of context-invariant lexical items in discourse-level assessment for PWA. Core lexicon checklists created for five word classes (nouns, verbs, adjectives, adverbs, and function words) were compared among the two criteria (frequency and percentage) using confirmatory factor analysis (CFA). To identify the source of measurement quality in core lexicon measures, several steps were taken. Discussion will focus on two main sources of variance that represent the quality of the measurement: factor loadings and residual variances. Results suggested that core lexicon measures based on both criteria reflect word retrieval ability in discourse and no differences in item contribution to the overall score emerged among the two. However, greater residual variances were found in percentage than in frequency, indicating that core lexicon lists based on percentage are more affected by measurement errors. Since the results of CFA demonstrated that frequency is a better criterion for the development of core lexicon measures, core lexicon items for the second aim of the study were identified by the frequency criterion. Consistent with expectations, function words showed an appropriate distribution of item difficulty and discrimination. Nouns and adjectives did not fit into the model. However, the possibility of potential development of universal lists for verbs and adverbs was found, despite the limited range of item difficulty.

Quality of measurement in core lexicon measures

The results of this study indicated that the different core lexicon measures constructed using two different approaches measure the same construct and are highly correlated with each other (See Figure 6.3 & 6.4). The magnitude of the relationship between core lexicon measures of the two different approaches was 0.98 for the Good Dog Carl story and 1.00 for the Picnic story. These high correlation estimates indicate that greater performance on the core lexicon list determined by frequency reflect higher scores on the core lexicon list developed by percentage for both story tasks. These results suggest that core

lexicon lists of these two approaches can be interchangeably used. It is plausible that for speakers to achieve a successful delivery of the story, they should rely on the discourse stimuli that constrain lexical use. The property of the discourse stimuli provided limited possibility of expansion for lexical use, which ultimately leads to the similar construct of core lexicon measures by different criteria. Moreover, core lexicon measures do not include the entire production of lexical items, and only contain a limited number of items. Due to the methodology of how the core lexicon measures are constructed, core lexicon checklists became a closed set of lexical items. Even if some speakers provide a rich vocabulary to illustrate details on pictures in which a majority of speakers did not pay attention, these lexical items are highly like to be eliminated through the computational process. Collectively, a combined effect of discourse stimuli and methodology of the measure accounts for the high covariance estimates between the two approaches.

With respect to the factor loadings of the observed variables (core lexicon checklists by word class), differences of their magnitude did not emerge by approaches. As a reminder, higher loadings of observed variables in one core lexicon measure over the other indicates that scores generated by the core lexicon measure are stronger indicators of the latent trait that the core lexicon measure conceptually purports to measure. That is, one of two measures having higher factor loadings is more accurate for testing word retrieval ability in the discourse produced by PWA that researchers and clinicians elect to use. The patterns of the factor loadings of core lexicon measures were very similar for both stories. Whereas core lexicon checklists based on the frequency criterion reflect the communalities of word retrieval ability in discourse greater than those based on the percentage criterion for content words, the core function word checklist based on the percentage criterion is a better measure than that based on the frequency criterion. The difference regarding the magnitude of the loadings by word class may stem from the size of the searching pool of lexical items to

produce. Specifically, a critical methodological point of using wordless picture books to elicit language samples in developing core lexicon measures is to acquire diverse vocabulary, which may draw more robust sampling for the measurement. Using the wordless picture books allowed for fully capitalizing on the strength of the discourse elicitation materials for acquiring various content words that have bigger boundaries of lexical items to be retrieved by speakers. However, since function words in nature consist of a finite number of items, even with the discourse stimuli, types of function words produced may be very limited. Moreover, studies investigating natural language use reported the limited use of function words in written or oral discourse of English speakers (Baayen et al., 1995; Chung & Pennebaker, 2007). Of the 100,000-word productive vocabulary that English speakers have, function words only account for less than 0.04% of this total (Baayen et al., 1995). Chung and Pennebaker identified commonly used function words from 950,000 text samples from 80,000 different people using various themes. The top 10 function words account for 20% of the words spoken. Possibly, function words having a relatively smaller pool of lexical items increased the likelihood of producing the same function words, which may lead to the results. Nonetheless, the point of emphasis is that no differences among the two criteria were found. Thus, the difference between content and function words will be further discussed in the following section.

The second factor to determine the quality of measurement is residual variances that generally indicate sources of measurement errors. The results regarding the residual variances were quite similar to the results from the factor loadings. Whereas core lexicon measures created based on the frequency criterion have less measurement errors for content words, those created based on the percentage criterion have less measurement errors for function words. Based on the current analysis, it is not obvious what drives the difference of measurement errors between function words and content words. It is possible that the limited

number of function words influence greater measurement errors for the frequency criterion. A plausible methodological concern to note about using the frequency criterion to identify the lexical items was that a speaker repeatedly produced one lexical item. If one speaker says the same item too many times, it could inflate the number of times that the item is being produced by all speakers, and ultimately has an impact on the construction of core lexicon measures. Although this concern has not been clearly addressed yet, it is reasonable to assume that the open set for lexical items in content words and the large sample size can offset the effects of repetition of certain lexical items. However, if one word class has a limited pool of lexical items, like function words, the order of what are the most frequently occurring function words in a lexical list can be greatly marred by the excessive repetition of lexical use. Indeed, due to the higher probability of producing the same function words compared to content words, the repetitive behavior cannot be systematically filtered out through the computational process of the development of core lexicon measures.

One might ask whether these measurement errors are serious enough to be considered in the quality of measurement. It is unfortunate that sources of measurement errors are difficult to determine. Mathematically, measurement error is regarded as the difference between a measured value and the true value of the construct, which indicates that measurement error is random. Kline (2010) noted that measurement errors are surrogate variables for all sources of residual variation unexplained through the model. In this case, the extent to which measurement errors influence the quality of measurement should be estimated in concert with the latent factor. Since model fit is adequately achieved, it can be concluded that measurement errors do not harm the measurement of core lexicon measures.

Overall, it seems that frequency is the better criterion to use for content words, but not for function words. Further, the differential magnitude of factor loadings and residual variances in function words and content words seems minor. Consequently, on the basis of

the higher factor loadings of the four word classes on the frequency criterion, even though the percentage criterion was found to be better for function words, using the percentage criterion for identifying lexical items for all types of word classes may be practical.

Universal Core lexicon lists

In addition to providing statistical guidance for a better criterion of core lexicon measures, the current study has evaluated the potential development and use of context-invariant core lexicon lists for clinical purposes. To do this, 272 PWA's narrative samples of the Cinderella story from AphasiaBank were used. First, we evaluated whether core lexicon production of the data met the assumptions of unidimensionality and local independence. Then, we examined overall fit of the 1-PL and 2-PL models. The analyses suggest that core lexicon checklists closely approximate the assumption of unidimensionality and local independence. Comparison of the overall fit of the 1-PL and 2-PL models suggests that the 2-PL IRT model mostly provided a better fit to the data (See Table 6.4). Thus, the 2-PL model was computed to estimate parameters of item difficulty and item discrimination. The analysis of the 2-PL model generates visual graphs that can be used to interpret the following item properties: Item characteristic curve, Item information curve, and Test Information function. The item characteristic curves (ICC) shows the relationship of measurement properties between person and items. The item information curve (IIC) shows the relationship between the latent trait and the probability of producing a correct response. In the core lexicon measure, producing a correct response denotes the production of the specific lexicon item. The Test Information Function (TIF) shows the relationship between the latent trait and precision of the overall test.

Function Words

The results of function word item calibration have provided initial evidence of how

core function word items influence the measurement of word retrieval ability in discourse produced by PWA with the possibility of clinical use of a universal function word core lexicon checklist. The final core lexicon checklist had 10 items, which can be a brief measure with test efficiency.

Based on a visual inspection of IIC, function word items can be roughly divided into four groups (See Figure 6.6-9). The first group, which includes item 4 (And) and 5 (Be), clearly provides the most information at low ability levels (between $\theta = -2$ and $\theta = -1$) but almost no information about high ability levels ($\theta > 0$). Likely, these items may be too simple for mild aphasia populations. Conversely, the second group includes item 11 (His) and 16 (On) that do not give much information overall but provide some information on those who are at high ability levels ($\theta > 2$). The third group includes item 2 (A), 9 (Her), 26 (They), and 27 (To), providing the most information at the ability parameter below 0. The fourth group includes item 6 (For) and 29 (With), and does not provide much information, but covers a wide range of ability levels. According to Baker's (2001) guidance for the interpretation of item discrimination parameters, items with greater than 1.70 discrimination parameter perfectly discriminate those who have high latent trait and low latent trait. Of the 10 final items, 9 items had discrimination parameters higher than 1.70. It therefore appears that core function word checklists are the best functioning measure of PWA' word retrieval ability at the discourse level.

The TIF for the function word items reveals that the core function words provide the most information for below average ability levels (about $\theta = -1$) but does not provide much information about high ability levels (above $\theta > 1$). This indicates that the checklist captures moderately and severely low levels of discourse-level word retrieval ability in PWA. It is best to interpret test information within this range. The characteristics of this measure can be also supported by the ICC. Many items included in the function word lists are generally difficult

to those who are at a low ability level, and few items are difficult to those who are above ability level 0. Four items (6, 11, 16, 29) can provide a different probability of producing that item at different ability levels of test takers ($\theta = 0, 1, 2$).

Overall, the finalized core function word checklist can adequately capture information from PWA whose latent trait lie approximately between -2 and +1 standard deviations of the mean of discourse-level word retrieval ability. However, interpreting information outside of the -2 and +1 range of ability is advisable with caution.

Verbs

Results from the item calibration for verbs were quite different from those for function words. A visual inspection of the IIC for each item shows that item 15 (Have) clearly has the most information about ability level around average ($\theta = 0$) or slightly below average (See Figure 6.10-13). Items 14 (Go), 13 (Get), and 26 (Leave) provide the next greatest information, while others barely provide information. In the ICC graph, many items have overlapping slopes, indicating that these items have different item discrimination values. According to Baker's (2010) guideline, of the 10 final items, only 3 items (Get, Go, Have) have item discrimination parameters greater than 1.70. Only 3 items (Take, make, Leave) have item discrimination parameters between 1.35 and 1.69, indicating that these items have high discrimination ability to separate those with high ability and low ability. The TIF for the verb items show that the finalized, universal core verb checklist provides the most information for average ability level ($\theta = 0$), and adequately capture information from PWA whose latent trait lie approximately between -1 and 2 standard deviations of the mean of discourse-level word retrieval ability.

Prior to the analysis, it was expected that the finalized core verb lists would be composed of light verbs that can be broadly used for various context. Following Berndt et al

(1997)' definition of light and heavy verbs, core verbs in the finalized list consist of 6 light verbs (Do, Get, Go, Have, Take, Make) and 3 heavy verbs (Think, Know, Leave). With respect to item difficulty and discrimination parameters of each verb item in the list, heavy verbs (Think, Know, Leave) are comparatively less discriminating and easy items compared to light verbs. It may be that these heavy verb items are easier for PWA due to the nature of the verb itself, and the difficulty with light verb retrieval is associated with higher levels of item difficulty and discrimination. Breedin and colleagues (1998) reported an opposite pattern that six of eight PWA with selective verb impairment used more semantically complex verbs (i.e., heavy verbs) than light verbs in the verb story completion task and narrative production. The researchers suggested that heavy verbs bearing more specific meaning are difficult to disrupt in their language system compared to light verbs bearing general, multiple meanings.

Another related, interesting finding is that based on the ICC, some items which can be categorized as heavy verbs (Know, Think) do not reach 100% probability of being produced by PWA who have great ability ($\theta = 4$). Based on the standard normal distribution, approximately the top 2% of aphasia participants, likely very mild aphasia, can be included in the population. Collectively, it is important to remember that no clear picture regarding the item properties by semantic weight arise, at least, in the discourse of the Cinderella story. If one was to investigate the use of light and heavy verbs using psychometric analysis, it could be explained with the comparison of aphasia subtypes (e.g., fluent versus non-fluent aphasia, or agrammatism versus non-agrammatsim), based on findings of specific verb impairments by the subgroups.

Additionally, it is noteworthy that item 15 (Have) is highly informative in those who have average or lower average ability of discourse-level word retrieval ability. According to Halliday's (1985) verb categories from functional perspectives, the item "Have" is

categorized as a relational verb, functioning as a main verb describing or evaluating a person or object in discourse. In an investigation regarding the semantic patterns of verb use in the discourse of PWA (Armstrong, 2001), four PWA produced less relational verbs compared to normal controls. Armstrong noted that PWA are impaired in the use of relational verbs in discourse, which reflects their reduced ability to name a target object. For example, the purpose of use of relational verbs is generally to describe background of the story. In the language samples from AphasiaBank, PWA used "Have" to explain the background of Cinderella having a stepmother and two stepsisters, or having animal friends, or not having anything to wear for going to the party. In the case that a speaker has difficulty accessing target lexicon item such as objects or people (e.g., Cinderella has …[step sisters]), they attempt to avoid producing these types of statements. Possibly, this item is an indicator of PWA's word retrieval ability at the discourse level, leading to the highest discrimination ability among the verb items.

Adverbs

In the finalized core lexicon list, 7 adverbs were included as universal items. Contrary to other word classes in the final list (i.e., function word, verb), all adverbs provide the most information in higher trait regions and seem more appropriate when the goal is to optimally measure trait levels at the $\theta = 1$ to 2 range (See Figure 6.14-17). Higher levels of information can be gained along the trait region associated with the range, while less information can be gained in regions below $\theta = 1$. This suggests that core adverbs in the final list are appropriate for assessing persons with aphasia with high ability levels of discourse production, and are not appropriate for assessing those with severe levels of aphasia.

Based on the standard normal distribution, approximately 13.5% of the participants with aphasia have a θ range between 1 and 2, and it is likely that moderate-to-mild and mild

aphasia can be included in that trait range. Moreover, the test information function exhibited that the finalized adverb lists has the highest peak of information, which achieves the highest level of precision at the $\theta = 2$, and lowest precision in the center and lower ability range.

Admittedly, the prime aspect to drive narrow measurement propensity in core adverb lists remains uncertain. In the field of second-language acquisition in English, adverb use has been regarded as an indicator of language proficiency (e.g., Grant & Ginther, 2000; Pérez-Paredes & Sánchez-Tornel, 2009) and lexical diversity (e.g., Lu, 2012). As such, one possible explanation for our finding may be that adverb emergence in discourse produced by PWA is associated with their comparatively higher function of lexical use or communicative ability. This speculation can be supported by Sarno and colleagues' (2005) study in which they provided comprehensive treatment for 18 PWA who are 3 to 12 month post-stroke. Unfortunately, the researchers did not fully disclose details on the treatment sessions implemented. It seems that the treatment consisted of group treatment, peer support, and conversational practice, etc. During and following the comprehensive language treatment, researchers tested PWA every three months using word fluency task by grammatical category, and found that the major qualitative change was found in the production of modifiers. They interpreted it as evidence of temporal patterns of word class expansion. Possibly, the finalized core adverb list is likely to be suitable for quantifying language gains of a patient who is mild aphasia during the treatment. However, adverbs cannot practically be categorized as a unitary category, having multiple sub-categorizations (e.g., locative adverbs, prepositional adverbs) with a distinct role or characteristics in utterances (Gilquin, 2007; Pérez-Paredes, Hernández, & Jiménez, 2011). In order to test this claim in the context of aphasic discourse, research pertaining to PWA's adverb production and their performance by different categories of adverbs should be explored.

CONCLUSIONS AND FUTURE DIRECTIONS: STUDY AIM 1

The current study contributes to the improvement of measurement quality of core lexicon measures by demonstrating a better criterion for developing core lexicon checklists. Comparisons between two criteria that have mainly been used in core lexicon studies confirmed that using frequency as a criterion may induce more accurate scoring and interpretation for content words, while the percentage criterion seems to be better suited for function words. However, that is not to say that researchers should continue to choose frequency as a criterion of constructing core lexicon measures in different discourse tasks. In future studies, the same approach should be replicated with other discourse tasks previously reported in core lexicon studies.

Additionally, the differential magnitude of factor loadings and residual variances across two criteria was the focus. At the same time, to avoid deviating from the main purpose of the current study, results regarding the validity of score interpretations from core lexicon checklists by five word classes were not specifically mentioned in the discussion. It is important to note, however, that the extent to which each core lexicon checklist explains word retrieval ability at the discourse level does not seem to be high for both stories. For example, the highest proportion of variance in the indicators that is associated with word retrieval ability is 54% when using the frequency criterion. Specifically, the proportion of variance in core verb checklists was 36% and 31% for GDC and Picnic, respectively, although core verb checklists were found to have clinical significance in assessing overall language performance in PWA (Kim et al., 2019). The strongest indicator of word retrieval ability in the core lexicon measures with the frequency criterion was interestingly adverb (~ 53%) for both stories. Additionally, the lowest residual variance of all indicators is 47% for both stories. This finding has methodological implication for this measure in that theoretically, there is room for improvement of the core lexicon measures by reducing the

irrelevant variance (i.e., residual variances). It is also important to note that as a clinical tool, it is not reasonable to draw inferences about the proportion of variance in the indicators using the current sample that consists of cognitively healthy adults. Future investigations should further determine validity of score interpretation of core lexicon measures using clinical populations (e.g., aphasia).

CONCLUSIONS AND FUTURE DIRECTIONS: STUDY AIM 2

Although successful with relatively large sample size, there are a few notable limitations of the current work. First, since no item banks for core lexicon measures exist, we used core lexicon items selected by two different narrative discourse tasks, which simulated item banks. Although using narrative discourse tasks with multiple pages of picture books is known to be reliable and sufficient to elicit diverse lexical items for normal and clinical populations, how closely the simulated item bank is to common item banks is not clear.

Moreover, the simulated core lexicon item bank was established by age, which leads to a number of missing values for certain lexicon items. In the case that a core lexicon item is included only in one age group, but not in other age groups, the item's responses in the age group are used for item calibrations, which led to many missing values in residual items prior to the final decision. As described in Figure 6.5, items having over 90% response in the missing value category were pruned at the beginning of the IRT process. Of 7 residual items in the noun category, only 4 items exhibited a good fit of the model but had more than 100 missing values. In computing the IRT modeling using R, missing values in the data set were treated as MAR (missing at random), which was the proper treatment for our data set. However, the more missing values a respondent has, the greater the size of standard error associated with the respondent's ability estimate. The more missing values for an item also will lead to greater standard error relevant to the item difficulty and discrimination.

Collectively, no core noun was left in the final list. Therefore, further research is needed to refine the measure with a larger data set for each age bin. This will make a separate item calibration by age group without missing data.

One potential concern about this finalized, universal core lexicon list is that measurement invariance (i.e., item invariance) was not considered and this must be investigated for clinical utilization. For example, how to construct the story using target words cannot be homogeneous across different ethnic groups. In the current study, a small proportion of minor ethnic groups (e.g., Asian, Hispanic) were included in the demographic composition of the sample, and the language samples obtained from the Cinderella story are known to be culturally entrenched. In this same vein, some lexical items are less discriminating for these underrepresented ethnic groups in the current data set compared to Whites. Within the same ethnic group, it is also possible that the finalized items do not have the same measurement properties depending on individuals' caring environments, especially for the Cinderella story. Therefore, repeating the same process could be useful in removing items that display cultural bias and adding items that are specific for the target group, which will in turn, enhance the measurement precision for core lexicon measures.

As many researchers state, there are different types of discourse elicitation tasks (e.g., storytelling, eventcasts, recounts) and different discourse genres (e.g., narratives, procedures, expository) imposing various cognitive and linguistic demands on speakers (Bliss & McCabe, 2006; Brady, Armstrong, & Mackenzie, 2005; Nicholas & Brookshire, 1993). It is also important to remember that the types of words being produced by speakers to deliver a message are associated with the core ideas of the discourse stimuli. Although some context-related items that are specifically linked to the story stimuli were attenuated through the course of analysis, the finalized core lexicon lists still include lexical items that are tightly connected with the Cinderella story, such as pronouns and heavy verbs. Therefore, the

application of the method with other types of discourse elicitation tasks remains to be investigated in future studies.

CLINICAL AND RESEARCH IMPLICATIONS

The current study provides the potential to improve the validity, precision, and efficiency of discourse performance in PWA by investigating item-level psychometrics of core lexicon measures. Although preliminary, the results of this study regarding the development of universal core lexicon list that can be globally used for various discourse tasks are also encouraging. First, it provides insight in the production of core lexicon items in discourse of aphasia patients. Core adverbs covered PWA at the high end of the ability scale, considering the most difficult items. Core function words may not be effective in measuring mild PWA because core function word list is considered an easy set. Because of this restricted measurement range of each checklist by word class, core lexicon measures can achieve greater measurement precision with a composite score of checklists. Second, it would seem likely that the finalized core function word is ideal in general use. Clearly, context exerts a significant effect on what words speakers say and pronoun use is the most related to the context. Although pronouns in the final core function word list might restrict the generalization of core lexicon checklists, it should be noted that pronouns revealed proper item difficulty and discrimination parameters in the current investigation and inclusion of pronouns in core lexicon lists in response to different characters in discourse stimuli is worth considering.

Potential focus on developing universal core lexicon measures is to offer practical solutions to clinicians in a way that day-to-day fluctuations in discourse performance of PWA are measured with many different discourse stimuli. A coherent understanding of discourse-level assessment in clinical settings is that while it is important to conjecture how PWA

perform their daily communicative tasks, it is quite unrealistic that clinicians administer discourse tasks due to practical reasons, such as a short time to elicit language samples and a lack of resources. In pursuit of growing demands for clinically manageable discourse tasks, core lexicon measures were initially introduced and have been investigated in recent years. By simplifying the process of assessment for clinicians, core lexicon measures have been viewed as an alternative means of capturing word retrieval ability in discourse produced by PWA. However, the limited number of core lexicon checklists previously developed and validated cannot be repeatedly used to monitor language changes in the same patients due to measurement issues. The best-case scenario may be to develop a set of multiple stimuli that are well-validated within the core lexicon framework. But, given the lengthy development and validation process of core lexicon checklists with a large data set which is the basic corpus for the core lexicon measures, this doesn't seem to be realistic. Therefore, a viable plan would repeat the item-level analysis for different discourse tasks. Once it is achieved, periodic assessment of discourse performance of PWA during the treatment can be conducted. In gaining a better understanding of patients' communicative skills on a regular basis, clinicians can adjust treatment plans, which ultimately improves the quality of practice.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Armstrong, L., Brady, M., Mackenzie, C., & Norrie, J. (2007). Transcription- less analysis of aphasic discourse: A clinician's dream or a possibility? *Aphasiology*, *21*(3–4), 355–374.

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics*, *32*(4), 346–359.

Bartels-Tobin, L. R., & Hinckley, J. J. (2005). Cognition and discourse production in right hemisphere disorder. *Journal of Neurolinguistics*, *18*(6), 461–477.

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkston, K. (2011). An introduction to item response theory and rasch models for speech-language pathologists. *American Journal of Speech-Language Pathology*, *20*, 243–259. https://doi.org/10.1044/1058-0360(2011/10-0079)

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Berndt, R. S., Haendiges, A. N., Mitchum, C. C., & Sandson, J. (1997). Verb retrieval in aphasia: 2. Relationship to sentence processing. *Brain and Language*, *56*, 107–137. https://doi.org/10.1006/brln.1997.1728

Bird, H., Howard, D., & Franklin, S. (2003). Verbs and nouns: The importance of being imageable. *Journal of Neurolinguistics*, *16*(2–3), 113–149.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.),Statistical theories of mental test scores (pp. 395–479). Reading, MA: Addison Wesley.

Bliss, L. S., & McCabe, A. (2006). Comparison of Discourse Genres: Clinical Implications. *Contemporary Issues in Communication Science and Disorders*, *33*(2), 126–137.

Bock, D. (1997). A brief history of item theory. *Educational measurement: issues and practice*, *16*(4), 21-33.

Boles, L., & Bombard, T. (1998). Conversational discourse analysis: Appropriate and useful sample sizes. *Aphasiology*, *12*(7–8), 547–560.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061.

Brady, M., Armstrong, L., & Mackenzie, C. (2005). Further evidence on topic use following right hemisphere brain damage: Procedural and descriptive discourse. *Aphasiology, 19*(8), 731–747. https://doi.org/10.1080/02687030500141430

Breedin, S. D., Saffran, E. M., & Schwartz, M. F. (1998). Semantic factors in verb retrieval: An effect of complexity. *Brain and Language, 63*(1), 1–31.

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample size and test-retest stability of connected speech measures for adults with aphasia. *Journal of Speech, Language, and Hearing Research, 37*(2), 399–407.

Brown, R. (1973). *A first language: The early stages.* Harvard U. Press.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Capilouto, G., Wright, H. H., & Wagovich, S. A. (2005). CIU and main event analyses of the structured discourse of older and younger adults. *Journal of Communication Disorders, 38*, 431–444. https://doi.org/10.1016/j.jcomdis.2005.03.005

Carragher, M., Sage, K., & Conroy, P. (2015). Preliminary analysis from a novel treatment targeting the exchange of new information within storytelling for people with nonfluent aphasia and their partners. *Aphasiology, 29*(11), 1383–1408.

Clark, E. V. (1978). Awareness of language: Some evidence from what children say and do. In *The child's conception of language* (pp. 17–43). Springer.

Coelho, C. A. (2002). Story narratives of adults with closed head injury and non-brain-injured adults: influence of socioeconomic status, elicitation task, and executive functioning. *Journal of Speech, Language, and Hearing Research : JSLHR*, *45*, 1232–1248. https://doi.org/10.1044/1092-4388(2002/099)

Coelho, C. A. (2007). Management of discourse deficits following traumatic brain injury: Progress, caveats, and needs. *Seminars in Speech and Language, 28*(02), 122–135. Copyright© 2007 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA.

Cook, K. F., O'Malley, K. J., & Roddey, T. S. (2005). Dynamic assessment of health outcomes: time to let the CAT out of the bag?. *Health services research, 40*(5p2), 1694-1711.

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*, *39*(11), 1125–1137. https://doi.org/10.1044/2015_AJSLP-14-0161

de Riesthal, M., & Diehl, S. K. (2018). Conceptual, methodological, and clinical considerations for a core outcome set for discourse. *Aphasiology, 32*(4), 469–471.

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: have we reached the tipping point? *Aphasiology*, *32*(4), 459–464.

Dillow, E. (2011). *Narrative Discourse in Aphasia: Main Concept and Core Lexicon Analyses of the Cinderella Story*. Columbia: University of South Carolina.

Doyle, P. J., Goda, A. J., & Spencer, K. A. (1995). The communicative informativeness and efficiency of connected discourse by adults with aphasia under structured and conversational sampling conditions. *American Journal of Speech-Language Pathology*, *4*(4), 130–134.

Elia, D., Liles, B. Z., Duffy, R. J., Coelho, C. A., & Belanger, S. A. (1994). An investigation of sample size in conversational analysis. *ASHA Convention, New Orleans, LA*.

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

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology*, *25*(11), 1414–1430. https://doi.org/10.1080/02687038.2011.603898

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

Finch, S., & Chater, N. (1992). Bootstrapping syntactic categories using statistical methods. *Background and Experiments in Machine Learning of Natural Language*, *229*, 235.

Fromm, D. A., Forbes, M., Holland, A., & MacWhinney, B. (2013). *PWAs and PBJs: Language for describing a simple procedure*.

Gauch, S., Wang, J., & Rachakonda, S. M. (1999). A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Transactions on Information Systems (TOIS)*, *17*(3), 250–269.

Gazella, J., & Stockman, I. J. (2003). Children's story retelling under different modality and task conditions: Implications for standardizing language sampling procedures. *American Journal of Speech-Language Pathology*, *12*, 61–72. https://doi.org/10.1044/1058-0360(2003/053)

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.

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

Gottron, T. (2009). Document word clouds: Visualising web documents as tag clouds to aid users in relevance decisions. *International Conference on Theory and Practice of Digital Libraries*, 94–105. Springer.

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

Hanlbleton, R. K., Swaminathan, H., & Rogers, D. J. (1991). *Fundamentals of item response theory*. SAGE Publications, Inc.

Hartley, L. L., & Jensen, P. J. (1991). Narrative and procedural discourse after closed head injury. *Brain Injury*, *5*(3), 267–285. https://doi.org/10.3109/02699059109008097

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55.

Ivanova, M. V., & Hallowell, B. (2013). A tutorial on aphasia test development in any language: Key substantive and psychometric considerations. *Aphasiology*, *27*(8), 891–920. https://doi.org/10.1080/02687038.2013.805728

Jesperson, O. (1965). *A modern English grammar based on historical principles*. London, Allen & Unwin.

Kertesz, A. (1982). *Western aphasia battery test manual*. Psychological Corp.

Kim, M., & Thompson, C. K. (2004). Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization. *Brain and Language*, *88*(1), 1–20.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284.

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

Kurland, J., & Stokes, P. (2018). Let's talk real talk: an argument to include conversation in a D-COS for aphasia research with an acknowledgment of the challenges ahead. *Aphasiology*, *32*(4), 475–478.

Law, S.-P., Kong, A. P.-H., & Lai, C. (2018). An analysis of topics and vocabulary in Chinese oral narratives by normal speakers and speakers with fluent aphasia. *Clinical Linguistics & Phonetics*, *32*(1), 88–99.

Lord, F. M., & Novick, M. R. (1968).Statistical theories of mental test scores. Reading, MA: Addison Wesley.

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 CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database*. MIT Press.

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., & Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiology*, *24*(6–8), 856–868. https://doi.org/10.1080/02687030903452632

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In *EVOLVING MODELS OF LANGUAGE* (pp. 58–71). https://doi.org/10.1191/0265532202lt221oa

Maouene, J., Laakso, A., & Smith, L. B. (2011). Object associations of early-learned light and heavy English verbs. *First Language*, *31*(1), 109–132.

Marshall, J., & Cairns, D. (2005). Therapy for sentence processing problems in aphasia: Working on thinking for speaking. *Aphasiology*, *19*(10/11), 1009–1020. https://doi.org/10.1080/02687030544000218

Mätzig, S., Druks, J., Masterson, J., & Vigliocco, G. (2009). Noun and verb differences in picture naming: Past studies and new evidence. *Cortex*, *45*(6), 738–758.

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

McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology*, *15*(10–11), 991–1006. https://doi.org/10.1080/02687040143000348

Meltzer-Asscher, A., & Thompson, C. K. (2014). The forgotten grammatical category: Adjective use in agrammatic aphasia. *Journal of Neurolinguistics*, *30*, 48–68.

Menn, L., Obler, L. K., & Miceli, G. (1990). *Agrammatic aphasia: A cross-language narrative sourcebook* (Vol. 2). John Benjamins Publishing.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383.

Miyamoto, T. (2000). *The Light Verb Construction in Japanese: the role of the verbal noun* (Vol. 29). John Benjamins Publishing.

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*(2), 338–350. https://doi.org/10.1044/jshr.3602.338

Nunnally, J., & Bernstein, I. (1994). Psychometric theory. McGraw Hill, New York

Oelschlaeger, M. L., & Thorne, J. C. (1999). Application of the correct information unit analysis to the naturally occurring conversation of a person with aphasia. *Journal of Speech, Language, and Hearing Research*, *42*(3), 636–648.

Olness, G. S., Gyger, J., & Thomas, K. (2012). Analysis of Narrative Functionality: Toward Evidence-based Approaches in Managed Care Settings. *Seminars in Speech and Language*, *33*, 55–67. https://doi.org/10.1055/s-0031-1301163

Olness, G. S., Ulatowska, H. K., Wertz, R. T., Thompson, J. L., & Auther, L. L. (2002). Discourse elicitation with pictorial stimuli in African Americans and Caucasians with and without aphasia. *Aphasiology*, *16*(4–6), 623–633. https://doi.org/10.1080/02687030244000095

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

Pinker, S. (2013). *Learnability and cognition: The acquisition of argument structure*. MIT press.

Pustejovsky, J., Anick, P., & Bergler, S. (1993). Lexical semantic techniques for corpus analysis. *Computational Linguistics*, *19*(2), 331–358.

R Core Team. (2019). R: *A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from the comprehensive R archive network (CRAN): https://www.R-project.org/,

Rider, J. D., Wright, H. H., Marshall, R. C., & Page, J. L. (2008). Using semantic feature analysis to improve contextual discourse in adults with aphasia. *American Journal of Speech-Language Pathology*.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1-36

Rubin, C., & Newton, E. (2003). *Capture the moment: The Pulitzer prize photographs*. Norton.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.Shapiro, L. P., & Levine, B. A. (1990). Verb processing during sentence comprehension in aphasia. *Brain and Language*, *38*(1), 21–47.

Thompson, C. K., Lange, K. L., Schneider, S. L., & Shapiro, L. P. (1997). Agrammatic and non-brain-damaged subjects' verb and verb argument structure production. *Aphasiology*, *11*(4–5), 473–490.

Tingley, E. C., Gleason, J. B., & 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. https://doi.org/10.1016/0021-9924(94)90038-8

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1–10.

Ulatowska, H. K., North, A. J., & Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain and Language*, *13*(2), 345–371. https://doi.org/https://doi.org/10.1016/0093-934X(81)90100-0

Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and 'theory of mind'in an agrammatic aphasic patient. *Current Biology*, *10*(12), 723–726.

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 standardisation of discourse measures? *Aphasiology*, *32*(4), 479–482.

Wallace, S. J., Worrall, L., Rose, T., & Le Dorze, G. (2014). Measuring outcomes in aphasia research: A review of current practice and an agenda for standardisation. *Aphasiology*, *28*(11), 1364–1384. https://doi.org/10.1080/02687038.2014.930262

Webster, J., Franklin, S., & Howard, D. (2007). An analysis of thematic and phrasal structure in people with aphasia: What more can we learn from the story of Cinderella? *Journal of Neurolinguistics*, *20*(5), 363–394.

Wechsler, D. (1945). A Standardized Memory Scale for Clinical Use. *The Journal of Psychology*, *19*, 87–95. https://doi.org/10.1080/00223980.1945.9917223

Weiss, J. A. (2012). *Differential Performance across Discourse Types in MCI and Dementia*. The Ohio State University.

Whitworth, A. (2018). The tipping point: are we nearly there yet? *Aphasiology*, *32*(4), 483–486.

Wright, H. H., & Capilouto, G. J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology*, *23*(10), 1295–1308. https://doi.org/10.1080/02687030902826844

Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, *8*, 484.

Table 6.1

Demographic and clinical characteristics of the participants

Characteristic	Participants (N = 272)	
	Women (N = 118)	Men (N = 154)
Ethnicity		
African American	15	14
Asian	0	3
Hispanic	4	2
Other	1	2
White	98	133
Education years completed		
M	15.11	15.74
SD	2.71	2.85
Minimum	11	8
Maximum	25	23
Age in years		
M	60.38	62.57
SD	13.9	11.6
Minimum	25.6	30.28
Maximum	90.7	85.72
Aphasia duration in years		
M	4.81	5.65
SD	4.25	5.09
Minimum	1	1
Maximum	20	30
WAB-R AQ		
M	76.19	70.16
SD	16.19	19.05
Minimum	34.5	10.8
Maximum	99.6	99.6

Note. WAB-R AQ = Western Aphasia Battery-Revised (Kertesz, 2006). Maximum WAB-R AQ raw score = 100.

Table 6.2

Summary of fit indices for configural, weak and strong invariance models (GDC)

	$\chi 2$	*df*	RMSEA[a] (90% CI)	SRMR[b]	CFI[c]	TLI[d]	**Δχ2**	**Δ***df*	**Δ**CFI
Configural Model	106.820**	29	0.076 (.061, .091)	0.034	0.973	0.957			
Weak Invariance Model	143.570**	33	0.084 (.071, .099)	0.061	0.961	0.947	36.751**	4	0.012
Modified Weak Invariance Model	110.596**	31	0.074 (.059, .089)	0.035	0.972	0.960	3.7762	2	0.001
Strong Invariance Model	628.294**	35	0.190 (.177, .203)	0.292	0.793	0.734	517.7**	4	0.179
Residual Invariance Model	512.975**	36	0.168 (.155, .181)	0.228	0.834	0.792	402.38**	5	0.138

Note. Changes in chi-squared test and CFI were compared among configural model, modified weak invariance, strong invariance, and residual invariance models.

[a]Root Mean Square Error of Approximation; [b]Standardized Root Mean Square Residual; [c]Comparative Fit Index; [d]Tucker-Lewis Index

*$p < .05$, **$p < .001$

Table 6.3

Summary of fit indices for configural, weak and strong invariance models (Picnic)

	$\chi 2$	*df*	RMSEA[a] (90% CI)	SRMR[b]	CFI[c]	TLI[d]	$\Delta\chi 2$	Δdf	ΔCFI
Configural Model	46.514*	29	0.036 (.014, .054)	0.025	0.992	0.988			
Weak Invariance Model	94.936**	33	0.063 (.049, .078)	0.052	0.973	0.963	48.421**	4	0.019
Modified Weak Invariance Model	47.521*	31	0.034 (.011, .052)	0.026	0.993	0.989	1.007	2	0.001
Strong Invariance Model	459.683**	35	0.161 (.148, .174)	0.244	0.813	0.760	412.16**	4	0.179
Residual Invariance Model	494.036**	36	0.165 (.152, .178)	0.203	0.798	0.778	446.51**	5	0.194

Note. Changes in chi-squared test and CFI were compared among configural model, modified weak invariance, strong invariance, and residual invariance models.

[a]Root Mean Square Error of Approximation; [b]Standardized Root Mean Square Residual; [c]Comparative Fit Index; [d]Tucker-Lewis Index

*$p < .05$, **$p < .001$

Table 6.4

Summary of model fit for core lexicon checklists by word class

Word class	Model	AIC[a]	BIC[b]	Log.Lik	LRT[c]	*df*	P value
Function words	1-PL	6845.34	6954.06	-3392.67			
	2-PL	6693.04	6903.23	-3288.52	208.3	28	<0.001
Verbs	1-PL	5806.36	5961.25	-2860.18			
	2-PL	5766.04	6068.61	-2799.02	122.33	44	<0.001
Adverbs	1-PL	6199.08	6353.97	-3056.54			
	2-PL	6183.07	6485.65	-3007.54	98.01	41	<0.001
Nouns	1-PL	3075.68	3205.35	-1501.84			
	2-PL	2737.28	2982.23	-1300.64	402.39	32	<0.001
Adjectives	1-PL	3808.89	3960.18	-1862.44			
	2-PL	3843.41	4138.79	-1839.71	45.47	40	0.255

[a]Akaike's Information Criterion; [b]Schwarz's Bayesian Information Criterion; [c]Likelihood Ratio Test

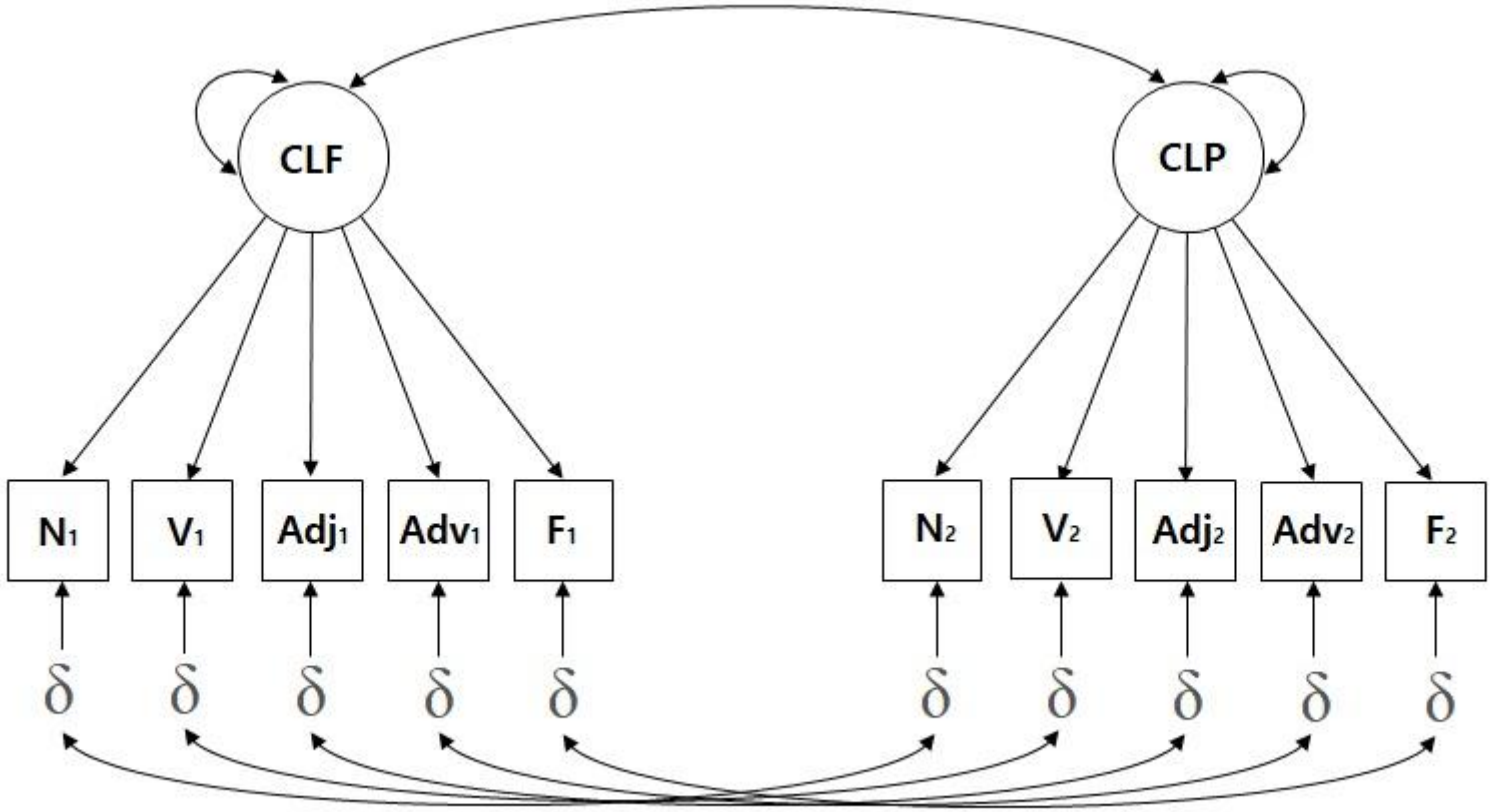

Figure 6.1. Path diagram

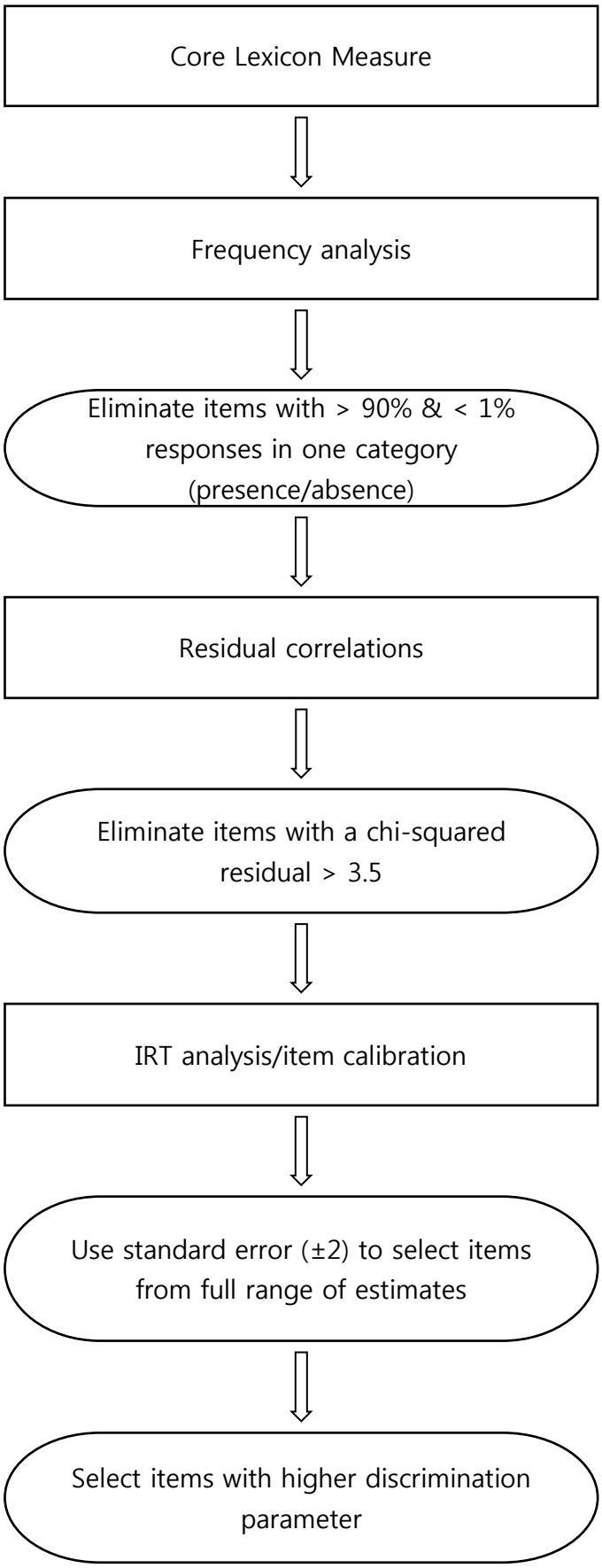

Figure 6.2. Flow chart for IRT analysis

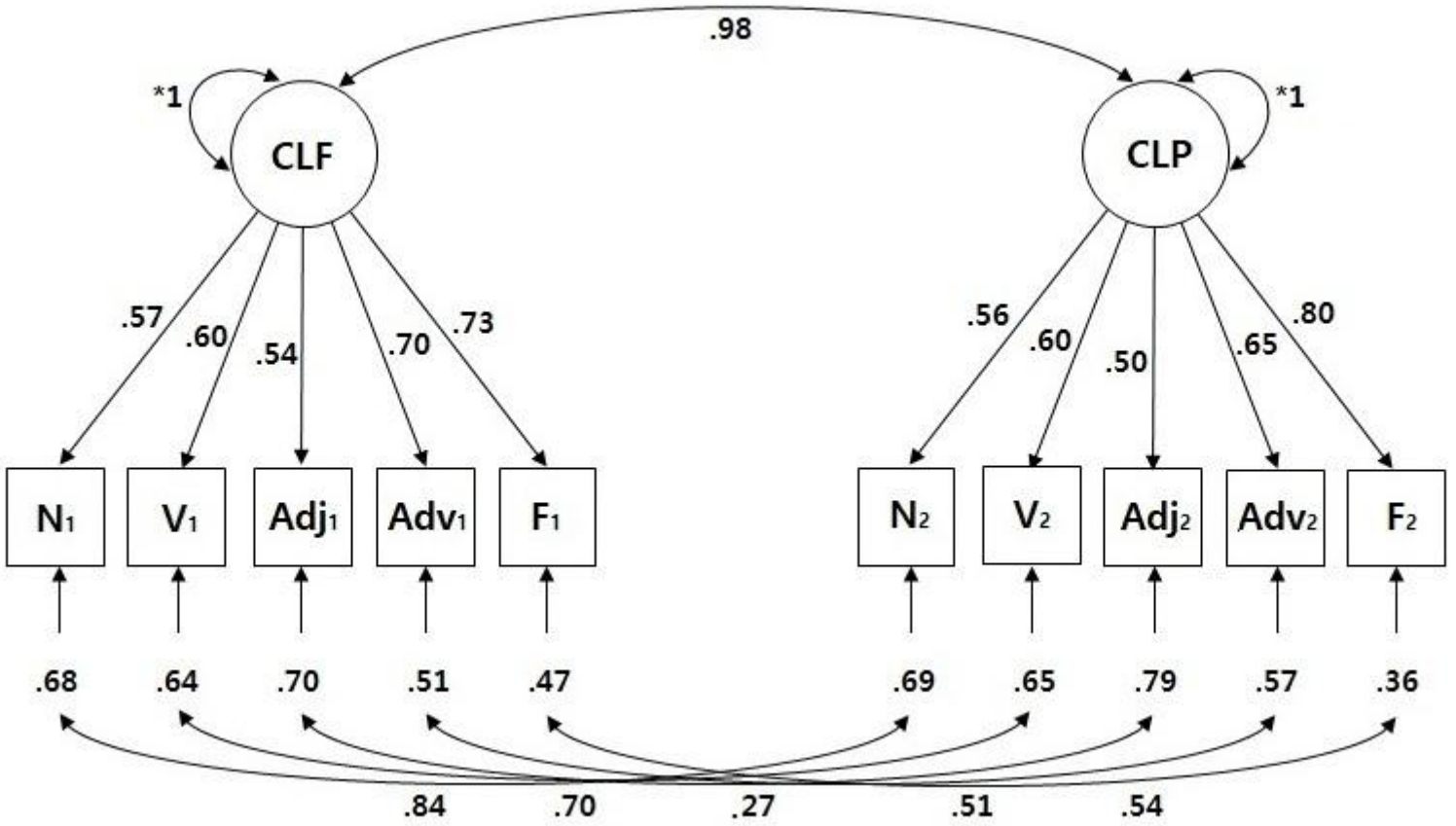

Figure 6.3. Standardized parameters for Good Dog Carl

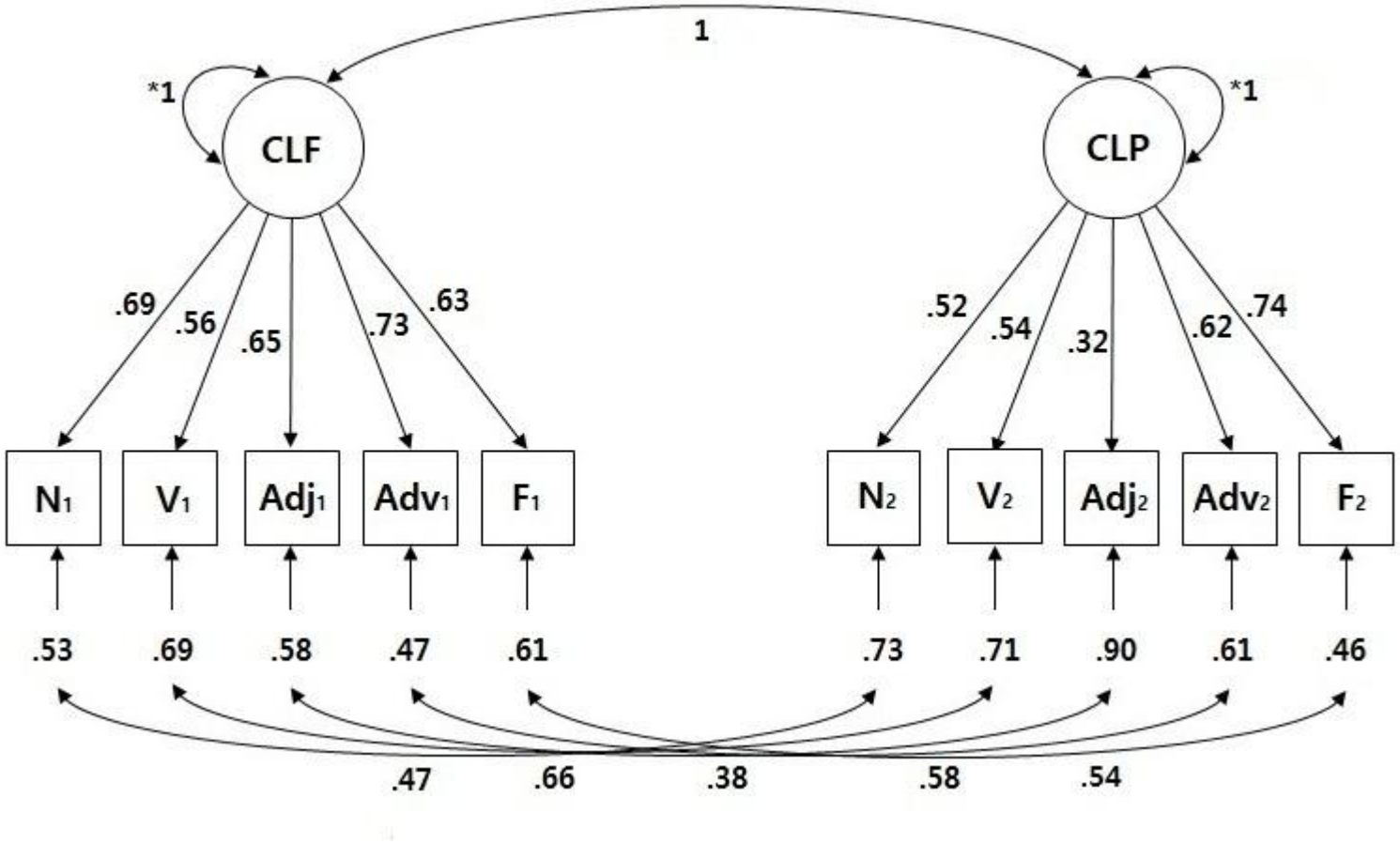

Figure 6.4. Standardized parameters for Picnic

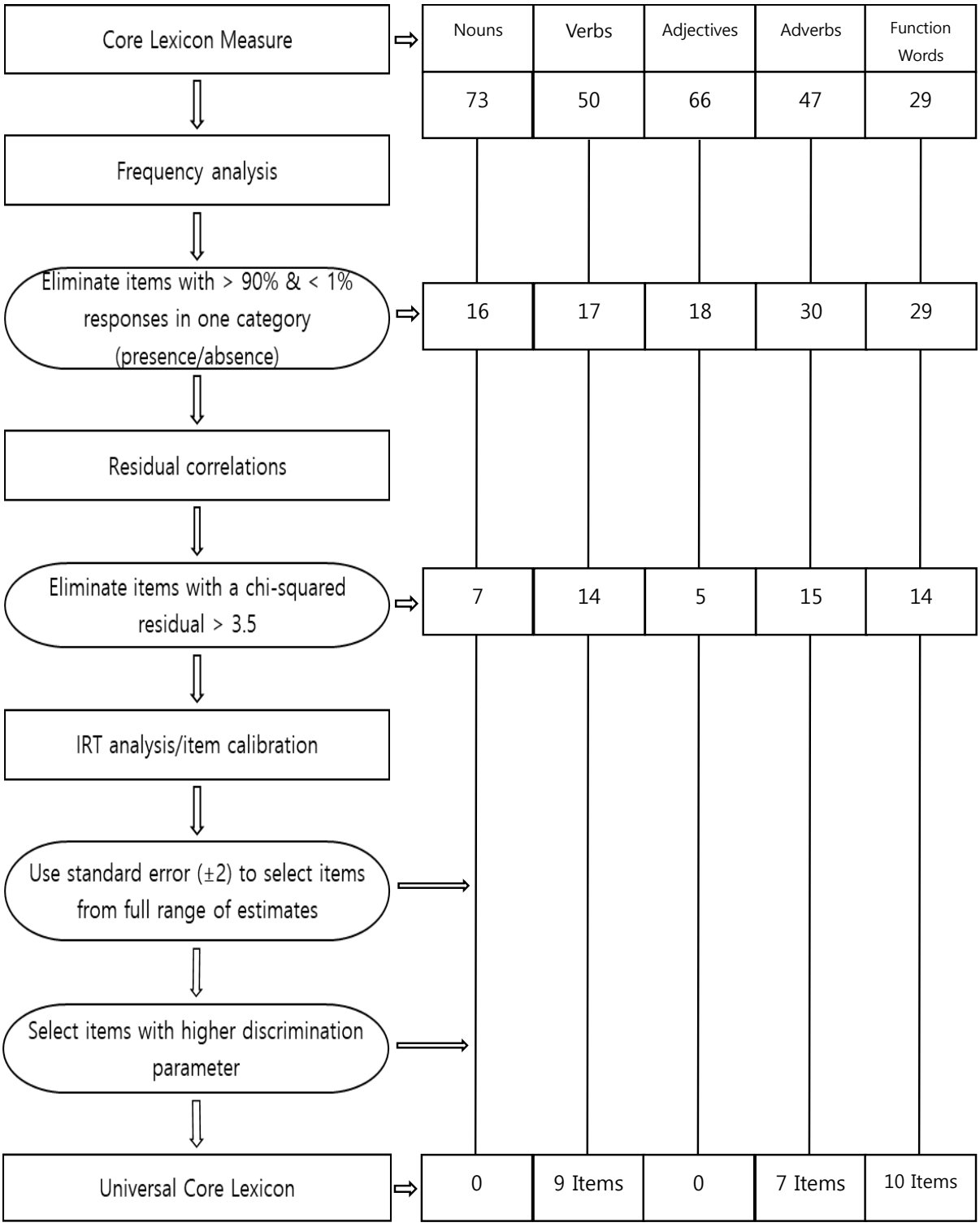

Figure 6.5. Flowchart of the procedures. The numbers to the right indicate the number of items remaining at the completion of previous step

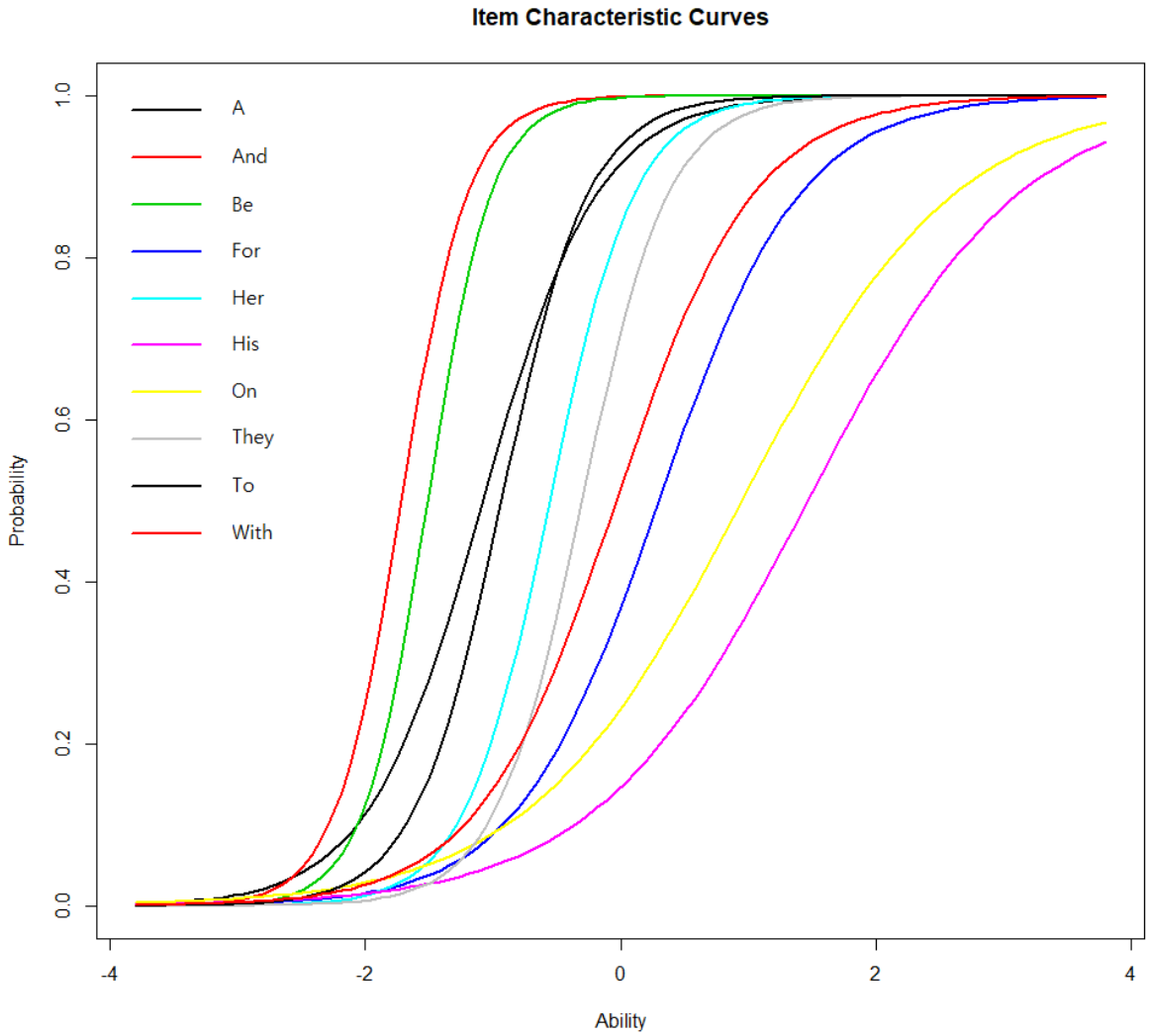

Figure 6.6. Item characteristic curves for function words

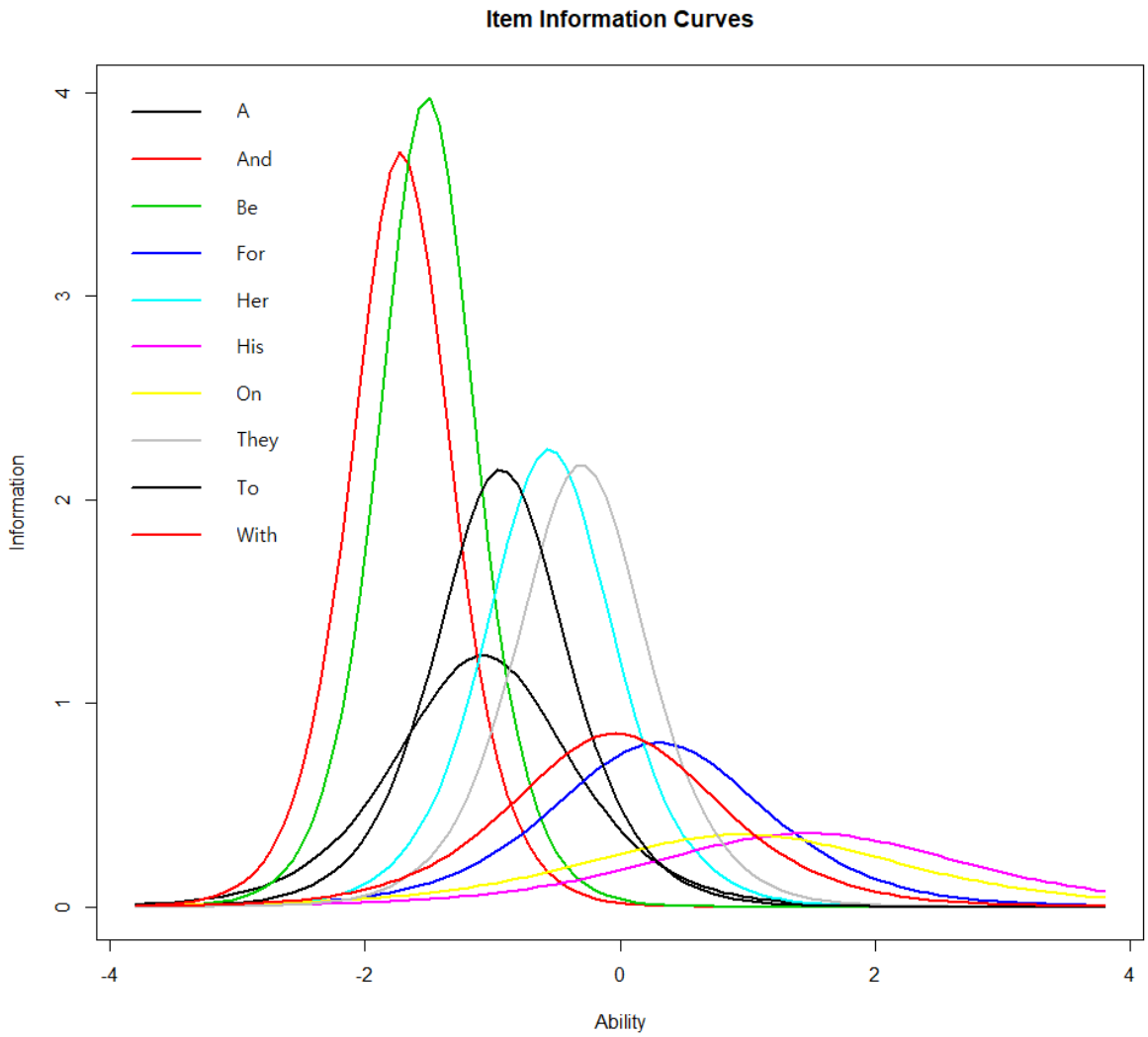

Figure 6.7. Item information curves for function words

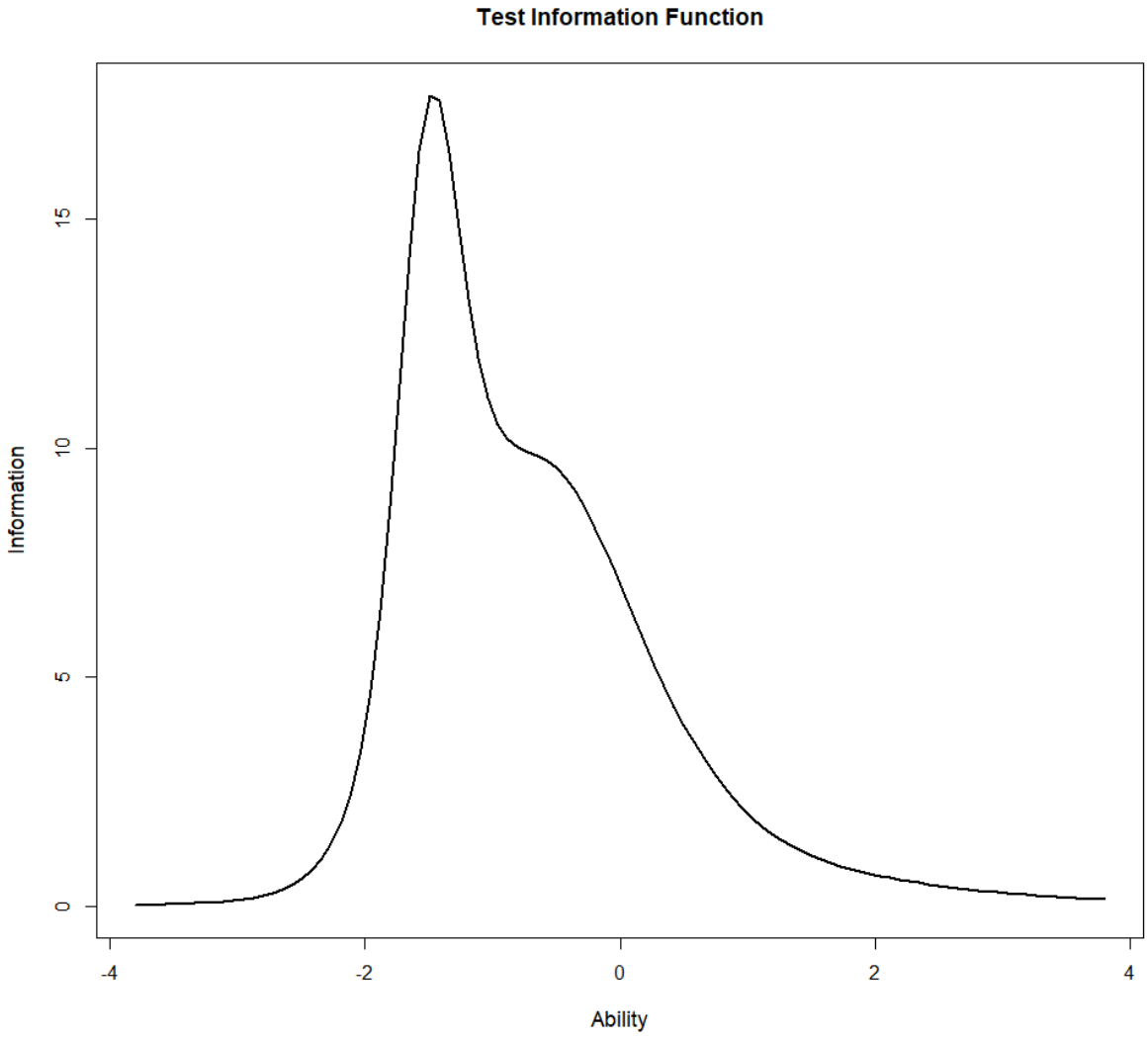

Figure 6.8. Test information function for function words

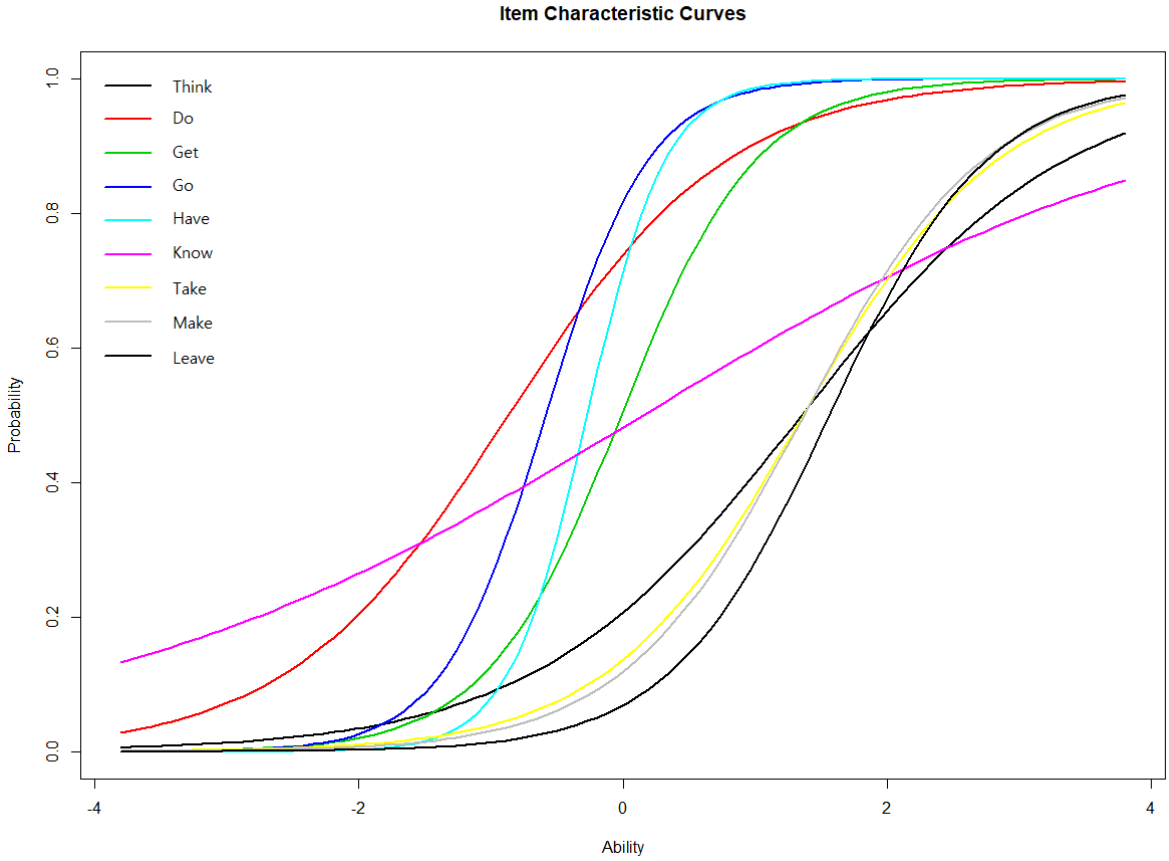

Figure 6.9. Item characteristic curves for verbs

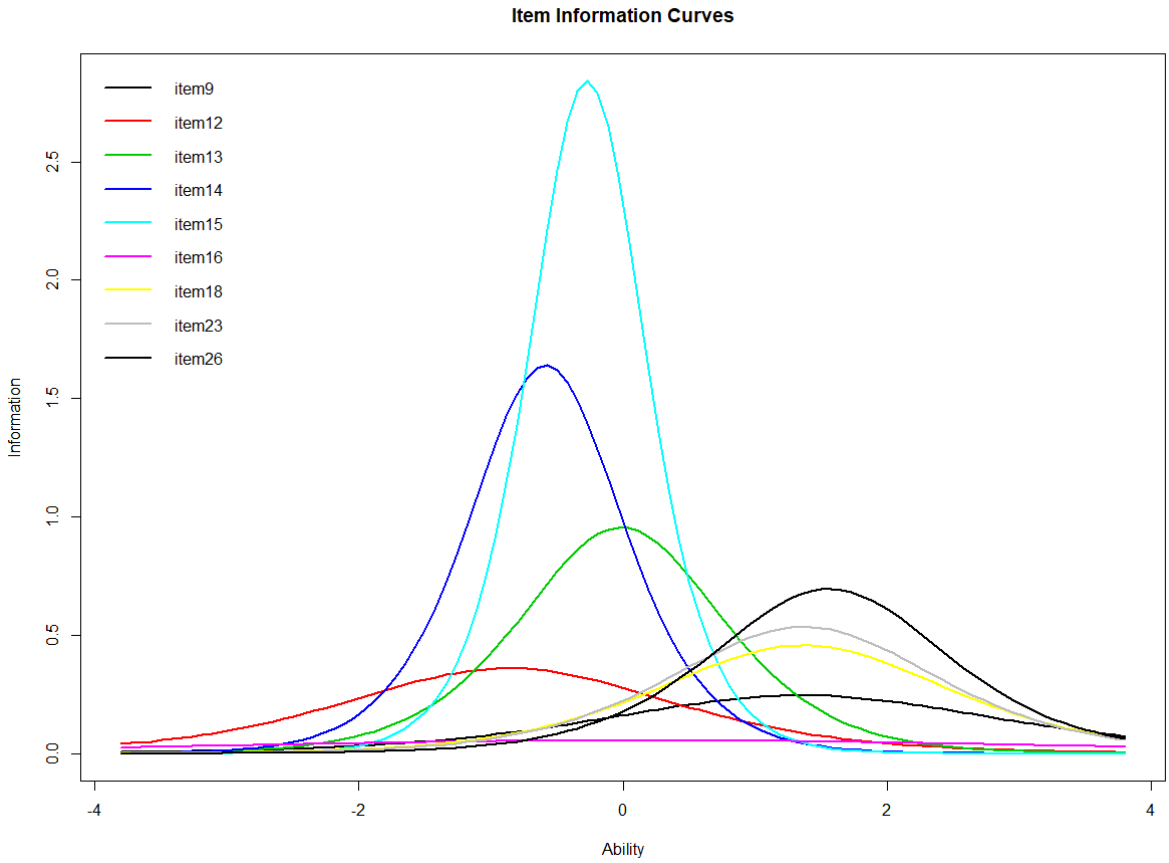

Figure 6.10. Item information curve for verbs

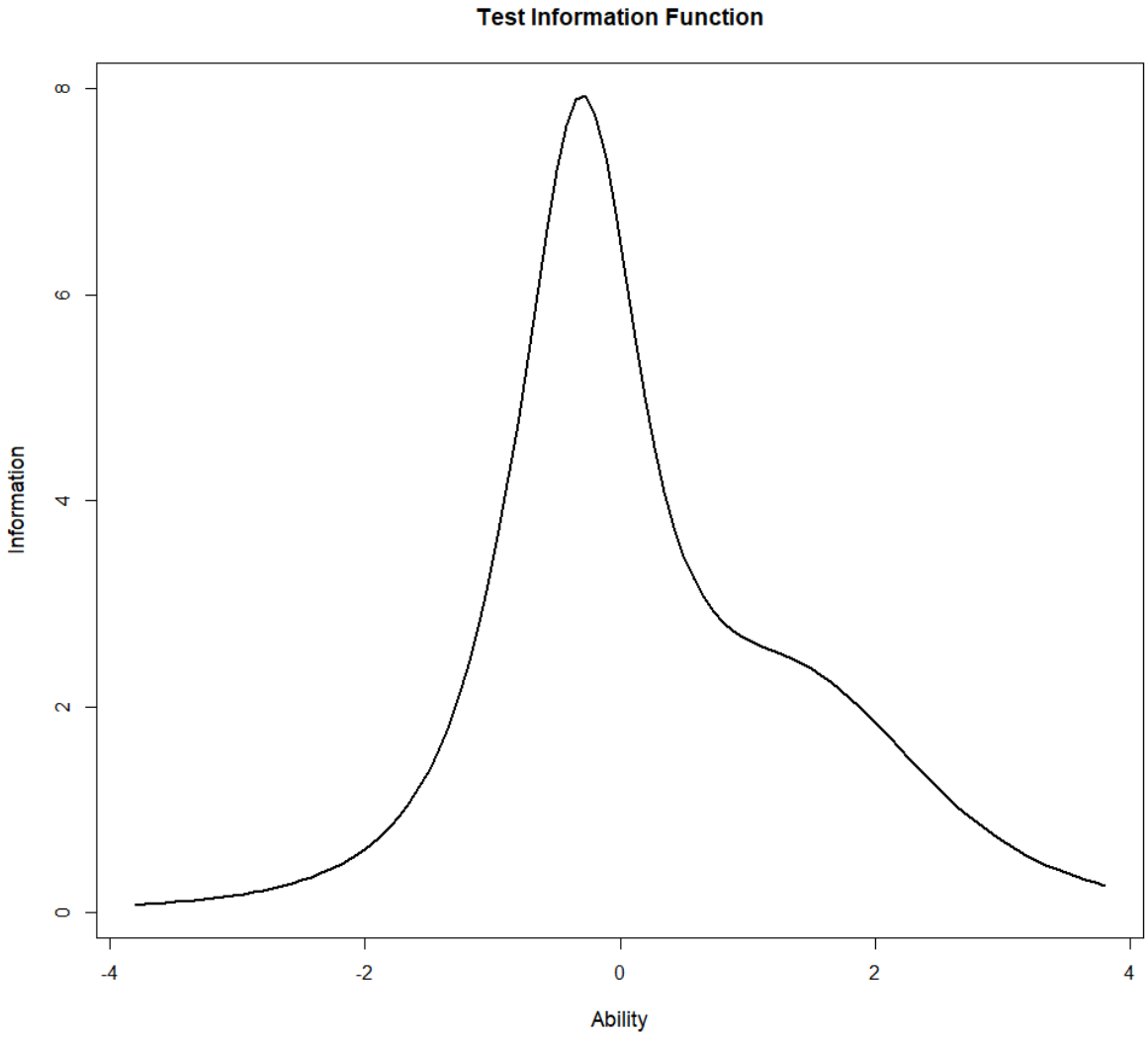

Figure 6.11. Test information function for verbs

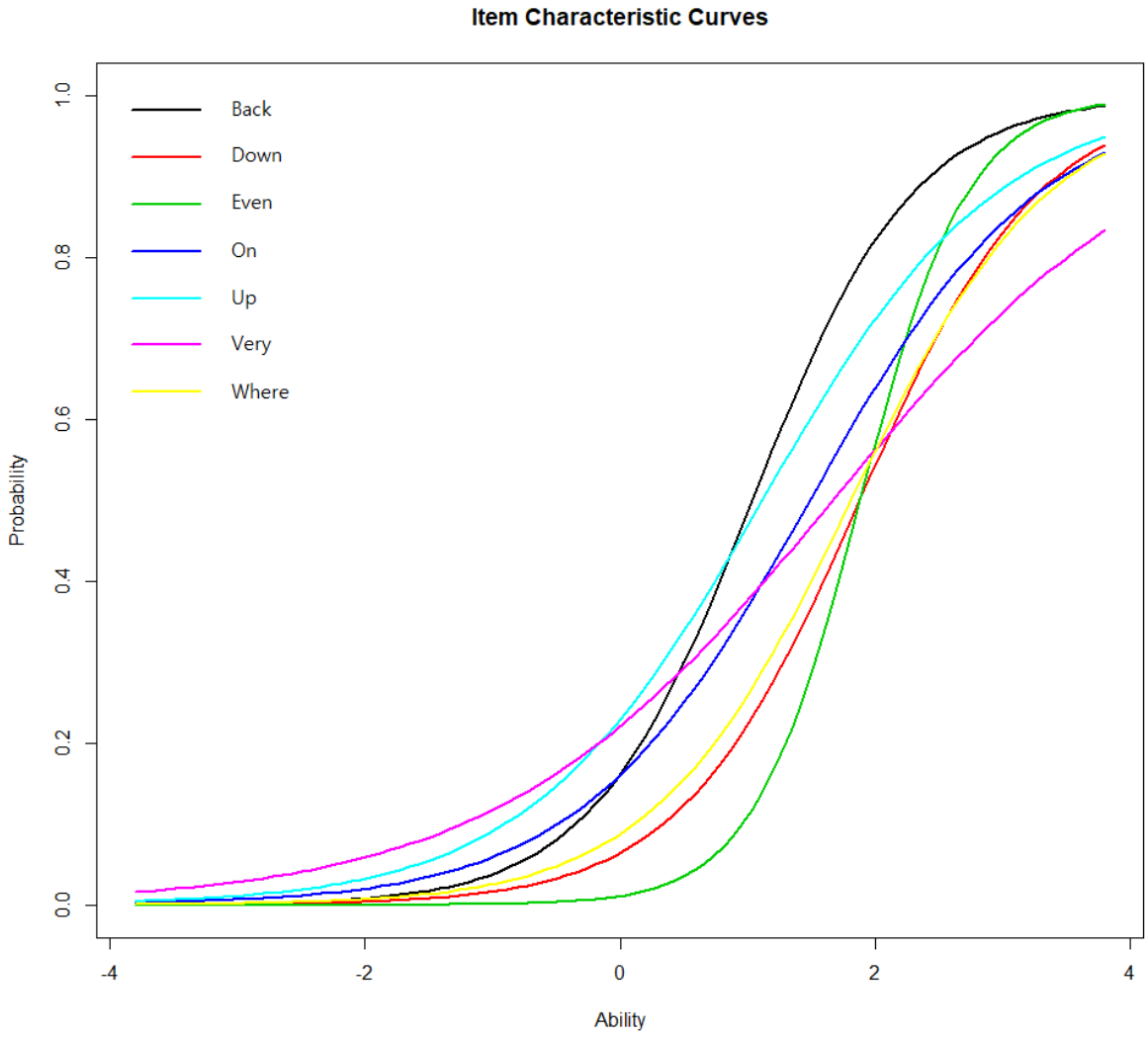

Figure 6.12. Item characteristic curves for adverbs

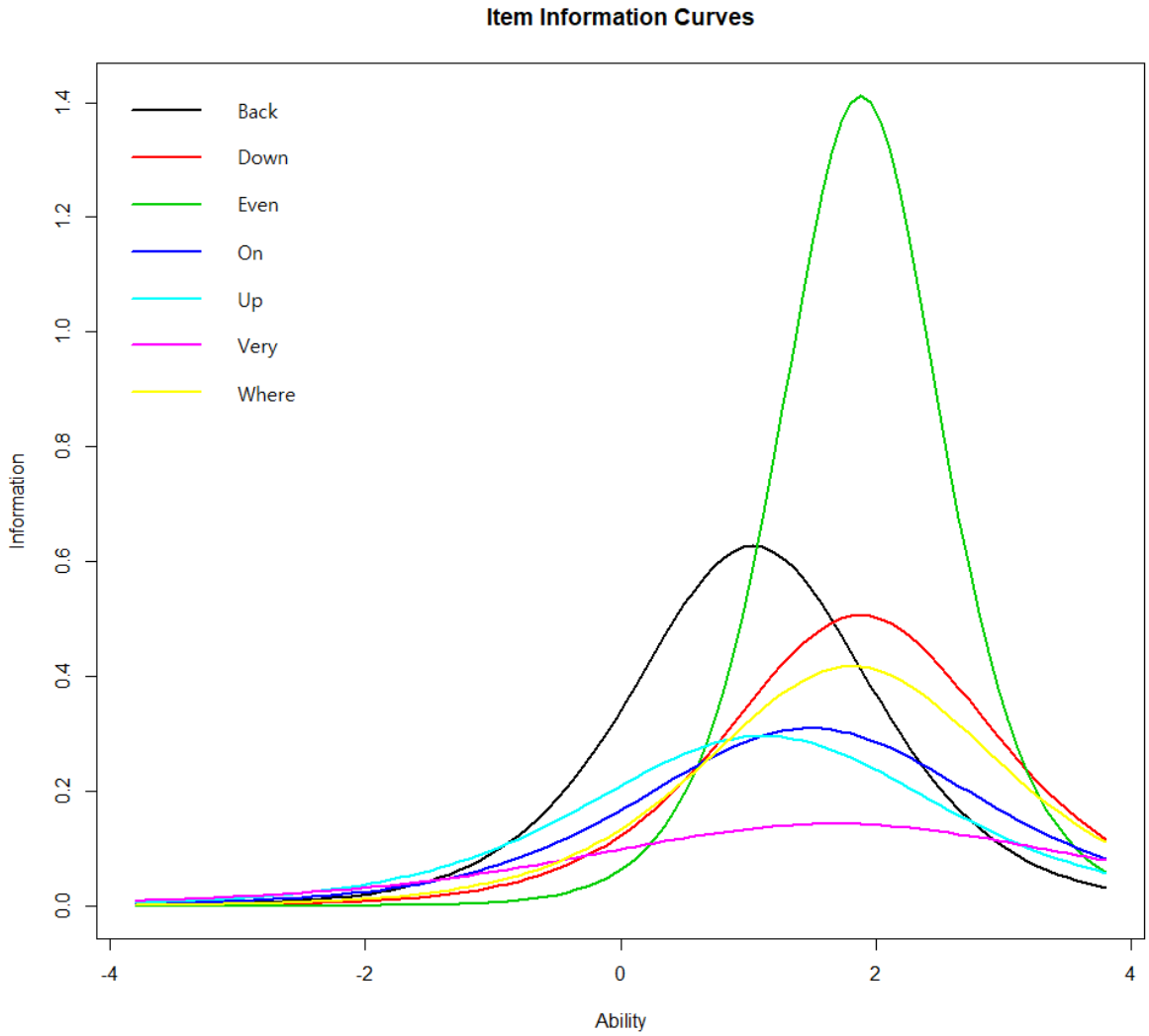

Figure 6.13. Item information curves for adverbs

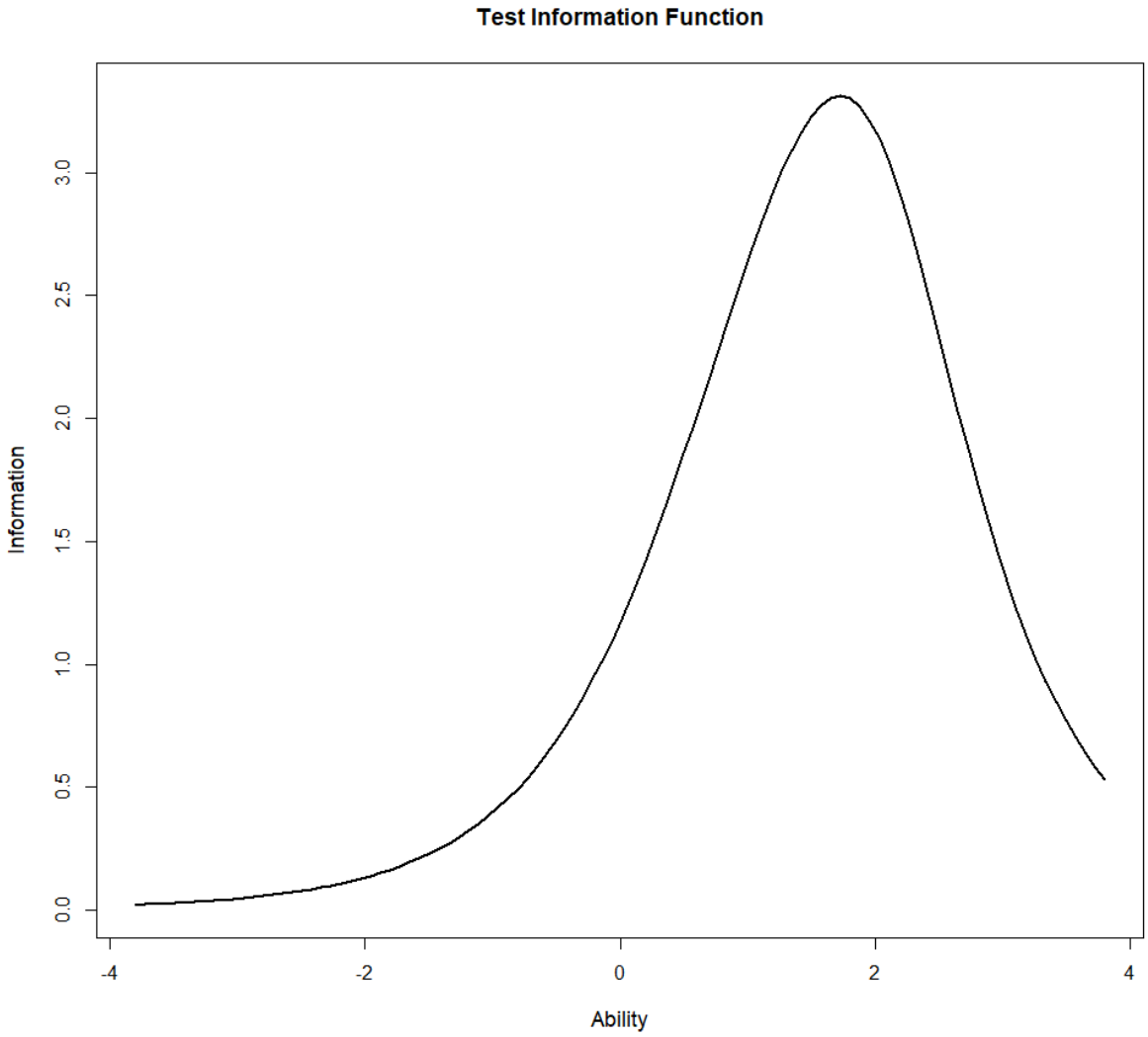

Figure 6.14. Test information function for adverbs

Appendix 6.A. Summary of final decision for universal core lexicon lists for five word classes

Nouns	χ^2	Difficulty	Discrimination	Missing Data	Inclusion/Exclusion	Notes
Day	0.00	1.88	1.30	163	Exclusion	High percentage of missing values
Mother	107.26	-0.05	0.81	None	Exclusion	Poor fit in the model
Mouse	54.25	1.49	1.20	None	Exclusion	Poor fit in the model
Thing	0.00	1.02	0.90	120	Exclusion	High percentage of missing values
Time	71.23	0.78	1.50	None	Exclusion	Poor fit in the model
Girl	0.00	1.63	1.96	197	Exclusion	High percentage of missing values
Place	0.00	1.65	0.81	194	Exclusion	High percentage of missing values

Note. Chi-square denotes an item fit index.

Verbs	χ^2	Difficulty	Discrimination	Missing Data	Inclusion/Exclusion	Notes
Think	0.00	1.28	1.05	120	Inclusion	
Do	8.17	-0.84	1.28	None	Inclusion	
Get	9.60	-0.04	2.06	None	Inclusion	
Go	16.38	-0.58	2.66	None	Inclusion	High item discrimination (with comparatively poor fit in the model)
Have	18.09	-0.28	3.74	None	Inclusion	High item discrimination (with comparatively poor fit in the model)
Know	2.24	0.12	0.51	None	Inclusion	
See	11.30	0.94	1.02	None	Exclusion	Poor fit in the model
Take	8.49	1.21	1.56	None	Inclusion	
Come	14.20	0.58	1.36	None	Exclusion	Poor fit in the model
Make	7.23	1.30	1.50	None	Inclusion	
Run	10.26	1.39	1.39	None	Exclusion	Poor fit in the model
Leave	0.00	1.56	1.59	151	Inclusion	
Say	27.84	0.29	1.49	None	Exclusion	Poor fit in the model
Find	10.67	0.61	1.94	None	Exclusion	Poor fit in the model

Note. Chi-square denotes an item fit index.

Adjectives	χ^2	Difficulty	Discrimination	Missing Data	Inclusion/Exclusion	Notes
Beautiful	29.14	1.75	0.93	None	Exclusion	Poor fit in the model
Big	110.28	1.65	1.02	None	Exclusion	Poor fit in the model
Good	8.76	1.94	0.47	None	Exclusion	Easy time with low item discrimination
Little	79.55	0.97	1.65	None	Exclusion	Poor fit in the model
Old	32.00	1.37	2.27	None	Exclusion	Poor fit in the model

Note. Chi-square denotes an item fit index.

Adverbs	χ^2	Difficulty	Discrimination	Missing Data	Inclusion/Exclusion	Notes
All	17.19	0.48	1.84	None	Exclusion	Poor fit in the model
Back	9.10	1.04	1.51	None	Inclusion	
Down	9.26	1.88	1.41	None	Inclusion	
Even	0.09	1.80	2.74	32	Inclusion	
Just	11.91	1.12	1.25	None	Exclusion	Poor fit in the model
Now	25.25	1.49	1.33	None	Exclusion	Poor fit in the model
On	7.35	1.48	1.10	None	Inclusion	
Out	11.85	1.12	1.25	None	Exclusion	Poor fit in the model
So	18.22	0.08	1.51	None	Exclusion	Poor fit in the model
Then	20.13	-0.21	1.16	None	Exclusion	Poor fit in the model
There	21.38	0.11	2.01	None	Exclusion	Poor fit in the model
Up	7.75	1.13	1.04	None	Inclusion	
Very	3.96	1.57	0.81	None	Inclusion	
Where	8.22	1.92	1.17	None	Inclusion	
Probably	0.00	1.53	14.96	165	Exclusion	High percentage of missing values

Note. Chi-square denotes an item fit index.

Function Words	χ^2	Difficulty	Discrimination	Missing Data	Inclusion/Exclusion	Notes
A	4.18	-1.03	2.30	None	Inclusion	
All	11.78	0.28	1.29	None	Exclusion	Poor fit in the model
And	4.80	-1.52	6.02	None	Inclusion	
Be	2.89	-1.43	4.05	None	Inclusion	
For	5.00	0.28	1.98	None	Inclusion	
Have	12.15	0.78	1.85	None	Exclusion	Poor fit in the model
Her	8.61	-0.52	3.28	None	Inclusion	
Him	11.84	1.42	1.68	None	Exclusion	Poor fit in the model
His	9.78	1.53	1.11	None	Inclusion	
On	9.60	0.93	1.20	None	Inclusion	
Them	15.97	1.01	1.35	None	Exclusion	Poor fit in the model
They	9.51	-0.29	2.92	None	Inclusion	
To	6.80	-0.87	3.29	None	Inclusion	
With	9.20	-0.03	2.10	None	Inclusion	

Note. Chi-square denotes an item fit index.

Appendix 6.B. Finalized universal core lexicon lists

Function Words	Verbs	Adverbs
A	Think	Back
And	Do	Down
Be	Get	Even
For	Go	On
Her	Have	Up
His	Know	Very
On	Take	Where
They	Make	
To	Leave	
With		