Long-term Stability, Test-Retest Reliability, and Lexical Properties of Core Nouns from Novel Discourse Stimuli

by

Qi Yukki Li

B.A., Rice University, 2016

M.S., University of Wisconsin–Madison, 2021

Submitted to the Graduate Faculty of the

School of Health and Rehabilitation Sciences in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF HEALTH AND REHABILITATION SCIENCES

This thesis was presented

by

Qi Yukki Li

It was defended on

February 23, 2024

and approved by

Michael Walsh Dickey, Ph.D., Professor, Communication Science and Disorders, University of Pittsburgh

Sarah E. Wallace, Ph.D., Professor, Communication Science and Disorders, University of Pittsburgh

Brielle C. Stark, Ph.D., Assistant Professor, Speech, Language and Hearing Sciences, Indiana University Bloomington

Thesis Advisor: William S. Evans, Ph.D., Associate Professor, Communication Science and Disorders, University of Pittsburgh

Stability, Test-Retest Reliability, and Lexical Properties of Core Nouns from Novel Discourse Stimuli

Qi Yukki Li, M.S.

University of Pittsburgh, 2024

Background: Core Lexicon Analysis (CLA) measures the use of keywords normed against the neurologically intact population during discourse. It is a promising tool to evaluate and monitor discourse production for people with aphasia (PWA), because it requires little transcription and training, and has been shown to have good interrater reliability and concurrent validity (Kim & Wright, 2020b). However, there are no studies on its test-retest reliability.

Aims: This study aims to investigate the test-retest reliability of CLA, specifically for core nouns, for a set of novel discourse stimuli. We also aimed to explore the effect of sample length on the test-retest reliability of CLA and characterize the phonological and semantic properties of the core nouns.

Methods & Procedures: Thirty-one non-brain damaged (NBD) community dwelling adults produced discourse in picture description and narrative storytelling tasks using novel visual stimuli. Fifteen participants saw discourse stimuli again after a mean of 149. Core noun lists, core noun scores, parameters related to test-retest reliability such as intraclass correlation coefficients (ICCs) and standard errors of measurement (SEMs), and sample lengths (defined as speaking durations) were obtained. Lexical properties of core nouns were extracted from corpus databases and normative studies (Balota et al., 2007; Fergadiotis et al., 2019; McRae et al., 2005; Johns, 2021).

Result: The ICCs for this set of stimuli ranged from –0.517 (95% Confidence Level [CI]: –0.875 to 0.033) to 0.695 (95% CI: [0.292, 0.887]), which are classified as "poor to moderate" according to Koo & Li (2016). Consistent with the ICCs, the group percent agreements ranged from 0 and 0.8, indicating considerable variability over time. Spearman's ρ between sample lengths and ICCs for individuals ranged from 0.14 to 0.72, indicating a small to moderate positive correlation between the two variables.

Conclusions: We conducted the first study on the test-retest reliability of CLA. Under our experimental conditions, the test-retest reliability of core nouns for picture description and story narrative tasks was poor, though it seems to increase slightly as discourse duration increases.

Table of Contents

List of Tables

List of Figures

Preface

"*The vitality of language lies in its ability to limn the actual, imagined and possible lives of its speakers, readers, writers. Although its poise is sometimes in displacing experience it is not a substitute for it. It arcs toward the place where meaning may lie."*

— Toni Morrison

Aphasia was the first condition that I learned about when I started to consider a path into speech language pathology. As a child, I found immense joy in the many worlds constructed by language. Reading and writing connected me deeply with actual and fictional characters and saved me from the constant loneliness that is inevitable in the life of an only child. Because of my love of language, I find aphasia a detestable force that can take away so much joy in life. I am grateful for this opportunity to contribute to aphasia research in working on this thesis project and thereby help those living with aphasia, however meager the help may seem. Meanwhile, as I further my journey into language and communication, I am comforted by the fact that, as quoted from the wise author Toni Morrison, we can still experience and cultivate meaning where language fails.

Thank you, Dr. Evans, for giving me an opportunity to work in your lab and learn about aphasia research and setting aside time to mentor me as I worked through this thesis. Thank you, Dr. Wallace, for modeling how to care and think deeply about others, to discern what is important and focus one's energy with purpose, and to live with generosity and authenticity. Thank you, Dr. Dickey, for sharing your wealth of knowledge and contagious passion, and doing all that you can to help me learn ever since I sat in your Transcription and Phonetics class as an English Language Learner. Thank you, Dr. Stark, for encouraging me since the conception of the project, and for inspiring me to always dig deeper, think more critically, and act more readily.

Besides my committee, I am fortunate to have Dr. Rob Cavanaugh and Alyssa Kelly as knowledgeable and wise mentors. Thank you for your patience and encouragement along the way. I am also grateful for Dr. Kendrea Garand, the moderator for my thesis defense in addition to bearing with me for 2.5 out of 5 semesters in my time at Pitt.

Last but not least, I wish to thank my family for supporting me through this huge career change and believing in my ability to excel. Your love is my shelter and strength.

1.0 Introduction

Aphasia is an acquired language disorder resulting from injury to certain cortical regions involved in language processing (Fridriksson et al., 2018; Kreisler et al., 2000; Naeser & Hayward, 1978). Any or all of the language functions can be affected in aphasia: auditory comprehension, speaking, reading comprehension, and writing. A cardinal symptom of aphasia is anomia, which refers to word-finding difficulties. Beyond anomia, there are diverse language ability profiles. Classification systems have been developed, but none can encapsulate the vast individual differences (Kasselimis et al., 2017).

The World Health Organization International Classification of Functioning, Disability, and Health framework (WHO ICF; 2001) provides a biopsychosocial perspective on health conditions, recognizing that health conditions not only cause physical discomfort, but also can affect mental health, limit access to resources, and affect family dynamics and social relationships. Based on the ICF framework, the Living with **A**phasia: **F**ramework for **O**utcome **M**easurement (A-FROM) model has been developed to encourage speech language pathologists (SLPs) to measure treatment effects beyond the improvements on linguistic skills (Kagan et al., 2008). In the A-FROM model, there are 4 domains related to living with aphasia: (1) personal identities, attitudes, and feelings; (2) severity of aphasia; (3) communication and language environment; and (4) participation in life situations. Two domains in the A-FROM model, severity of aphasia and participation in life situation, are directly affected by a person's ability to understand and produce discourse, while the domain of personal identities, attitudes and feelings can be indirectly affected by it. Improving discourse can bring about positive changes in multiple domains of the life of PWA. Therefore, measuring discourse skills should be an essential component in evaluating the impact of aphasia

and should be regularly conducted to assess the effect of aphasia treatment on meaningful change in everyday communication.

1.1 Discourse Characteristics in Aphasia

Conventionally, discourse is defined as language more complex than a single sentence (i.e., connected speech). Linnik et al. (2016) composed this elegant definition: "a flow of information put into words, organized in order to meet specific communication goals, and shaped by situational factors" (p.3). This definition illustrates the multilevel nature of discourse: words, the organization around the central idea or communication goal, and the situation shaping the specific language used. To comprehensively evaluate the impact of aphasia using the A-FROM model, one must consider how discourse is affected in aphasia, and one's consideration must include multiple components or levels of discourse. Dipper et al. (2021) synthesized 4 models and 6 frameworks of discourse to propose the following framework, **L**inguistic **U**nderpinning of **N**arratives in **A**phasia (LUNA), as a tool to describe how discourse can be analyzed in aphasia (figure 1).

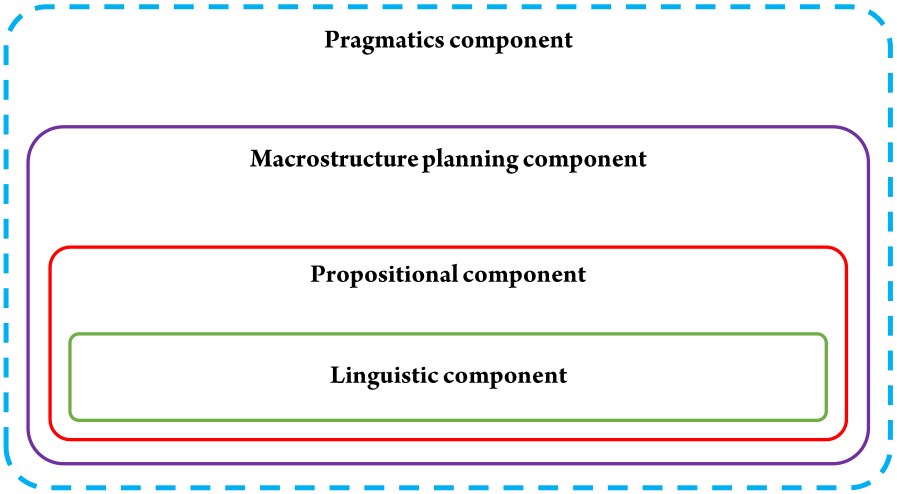

Figure 1 the LUNA framework (Dipper et al., 2021). Reproduced here with the copyright defined in the Creative Commons License (CC BY).

There are 4 interacting components of discourse in the LUNA framework. The first component, *pragmatics*, refers to the environmental and interpersonal factors that affect discourse, where a speaker determines the complexity and formality of language to use, how much background information to include, so that their discourse is appropriate for the occasion and interesting to the listener(s). For PWA, pragmatics may be affected by the inability to retrieve the appropriate lexical form of a concept that conveys too little information or is too casual for the situation (Linnik et al., 2022). It can also be affected by co-occurring attention and short-term memory deficits that lead to poor topic maintenance (Cahana-Amitay & Jenkins, 2018; Henderson et al., 2017). Next, the *macrostructure planning* component refers to the organization of discourse that contains the structural elements that befit the intended genre. For example, when telling a personal story, a speaker includes information on the setting, characters, and usually tells the story in chronological order. Meanwhile, when giving a presentation, a speaker typically describes logical relationships instead of temporal ones. For PWA, macrostructure may be affected by co-

occurring memory deficits that make crucial details inaccessible and co-occurring executive function deficits that make planning more challenging (Cannizzaro & Coelho, 2013). Next, the *propositional* component refers to the utterances that embody the macrostructure, such as those setting the stage for a story, signaling transitions between one theme and another, and clarifying logical relationships. For PWA, proposition may be affected by underutilization of verbs or function words signaling temporal or logical relationships (Bryant et al., 2013). Finally, the *linguistic* component refers to the actual language produced to actualize a discourse, including language domains such as syntax, semantics, morphology, and phonology. It is on this component that the impact of aphasia is most salient (Andreetta et al., 2012). Besides anomia, semantic or phonological paraphasias, which are words with different meanings ("table" for "clock") or sounds ("tok" for "clock") erroneously produced, are often present in discourse produced by PWA, hindering accurate transmission of information and feelings.

It is important to note that the potential areas of discourse impairments mentioned above have not been consistently demonstrated when researchers compared the performances of NBD participants and PWA (Linnik et al., 2016). The reasons proposed are the following: heterogeneity in multiple areas including conceptual definitions, discourse genre studied, and the measures used for underlying constructs; limitations in study design such as small sample sizes; and insufficient understanding of the relationship between micro- and macrolinguistic levels of discourse. To ensure that the outcomes of discourse studies are comparable and trustworthy, we must develop and use discourse measurements that are clearly defined, valid, and reliable.

1.2 Discourse Measures in Aphasiology

Bryant et al. (2016) performed a literature search and found that there were over 500 different measures for discourse used, spanning 12 different categories in 3 broad clusters: language productivity, information content, and grammatical complexity. These measures can be mapped onto the LUNA framework as shown in figure 2.

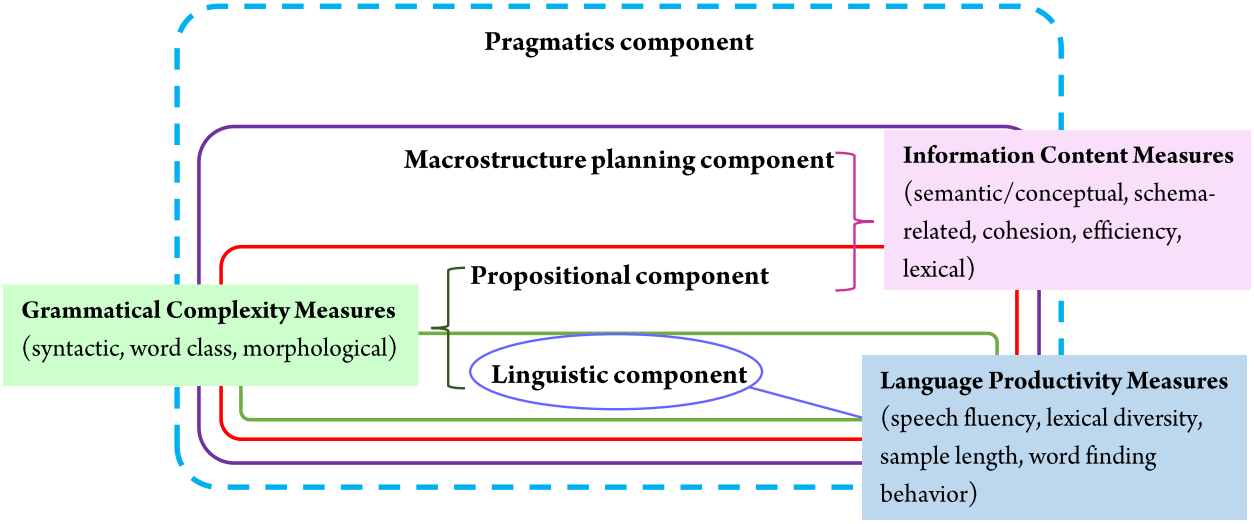

Figure 2 Existing discourse measures mapped onto the LUNA framework

The sheer number and heterogeneity in definitions pose a formidable challenge to achieve high quality evaluation and monitoring of progress in discourse for PWA, and many aphasiologists recognize the need for standardizing discourse assessment. Dietz and Boyle (2017) made a compelling case for developing a Core Outcome Set (COS), a set of clearly defined measures that are minimally required when assessing discourse. In 2019, a working group, **Fo**stering **Qu**ality of **S**poken Discourse in **Aphasia** (FOQUSAphasia; foqusaphasia.com), was established with two initial task forces: identifying best practices for reporting spoken discourse outcomes and improving data quality for studies on spoken discourse (Stark et al., 2021). The first task force

culminated in a set of best practices in reporting data on spoken discourse resulted from 3 rounds of electronic Delphi surveys of experts (Stark et al., 2023a); the checklist completed for this thesis is included in Appendix A.

Developing a COS in discourse requires critical evaluation of existing measures. Only a handful of discourse measures have been found to have robust psychometric properties. One example of these measures is correct information units (CIUs). Nicholas and Brookshire (1993) calculated that the interjudge reliability and intrajudge reliability of CIUs were over 90% and 95% respectively. In the same study, the authors also found percent CIUs to be a stable measure over three sessions. However, the interjudge reliability of CIUs obtained in Kendall et al. (2008)'s treatment study was 0.70. Pritchard et al. (2018) reviewed the measures used to analyze narrative, expositional, and personal discourse tasks, and concluded that Story Grammar, Topic Coherence, Reference Chains, and Predicate Argument Structures met criteria scores for acceptability, validity and reliability (the interjudge reliability ranged from 90% to 95%, while the intrajudge reliability ranged from 92% and 97%), suggesting that these measures have strong psychometric properties.

In addition to the high number of different discourse measures, the type of discourse tasks affects the complexity and content of language elicited. Specifically, microlinguistic properties have been found to correlate with discourse task types. For example, Stark (2019) analyzed language samples of three different discourse types, narratives (such as story retell tasks), expositional (such as picture description tasks) and procedural (such as describing how to make a peanut butter and jelly sandwich), and found that certain microlinguistic measures, namely mean length of utterance (MLU), propositional density, verbs per utterance, noun-verb ratio, and tokens are significantly correlated with discourse genre for NBD participants. Of the tasks evaluated, both people with and without aphasia tend to produce the highest propositional density for the story

retell narrative task (i.e., telling the Cinderella story). Propositional density (PD) is a measure of the average number of ideas per word (Fromm et al., 2016); it was calculated by summing the number of adjectives, adverbs, conjunctions, and prepositions and dividing this sum by the total number of words. By this definition, a language sample with a higher PD likely has a richer content. This finding suggests that clinicians and researchers should specify the discourse tasks used when reporting the outcomes and that narratives may capture more propositional ideas, and therefore are richer in content than the other types of discourse.

Clinically, objective discourse analysis, i.e., using objectively defined measures to analyze a discourse sample, is infrequently conducted. Most objective discourse measures require orthographic transcription of language samples. Bryant et al. (2017) conducted a survey for SLPs from several countries including the United States and Australia, and found that only around a quarter of the respondents included detailed transcription-based analyses for discourse samples. In contrast, half of the respondents agreed with the statement that "detailed linguistic analysis is important for the assessment of language in aphasia", indicating most clinicians view discourse as an important reflection of a person's communication abilities. The time demand for transcription, transcription service availability and need for training on discourse analysis, as well as perceived clinical values (i.e., whether the information gathered from discourse analysis is necessary to plan treatment for a specific client) are the reported limitations. The mismatch between attitudes towards objective discourse analysis and its current clinical application suggests that the time, training, and resources needed for transcription is a significant barrier for clinical assessment of discourse skills.

1.3 Core Lexicon Analysis

Discourse measures that require little or no transcription are more likely to be clinically utilized. **C**ore **L**exicon **A**nalysis (CLA) is one such time-efficient measure. Core lexicons are defined as "the pivotal lexical items required to produce a semantically meaningful and coherent narrative" (Kim et al., 2019, p.64). It was conceived to measure the word retrieval ability at the discourse level. The concept of CLA was first suggested by MacWhinney et al. in 2010, though the term was not coined until 2013 by Hudespeth et al. It was first applied to a story telling task using the Cinderella wordless picture book, where MacWhinney and colleagues conducted lexical frequency analysis using **C**omputerized **L**anguage **An**alysis (CLAN), and generated 10 most frequently used nouns and 10 most frequently used verbs by PWA and people without aphasia. They noted the percentage of nonspecific nouns or light verbs, which were verbs with less semantic complexity, were higher in the narratives produced by PWA. Since then, there have been numerous studies on CLA, including development of core lexicon checklists for different types of narrative elicitation tasks (Dalton et al., 2020). Recent studies on CLA have suggested its value in clinical use, including sensitivity to aphasia subtypes (Dalton & Richardson, 2015), mild aphasia (Fromm et al., 2017), and recovery from acute to chronic stages (Kim et al., 2023).

Several psychometric properties of CLA have been established. Kim and Wright (2020b) demonstrated that CLA of two wordless picture books, *Good Dog, Carl* (*GDC*; Day, 1985), and *Picnic* (McCully, 1984), yielded excellent interrater reliability and good concurrent validity with two macrolinguistic measures, coherence and thematic units. Coherence is a measure of how relevant a linguistic unit is to the main topic, and thematic units, a normative measure of information related to the main topic (Linnk et al., 2022; Marini et al., 2015). The finding that there was concurrent validity between CLA and these two macrolinguistic measures suggests that

CLA not only reflects an individual's word retrieval ability, but also the informativeness of their lexical decisions. In terms of microlinguistic measures, core nouns from *Picnic* were found to significantly correlate with syntactic complexity and lexical diversity, while those from *GDC* did not. The authors speculated that the stronger correlations between the core nouns in *Picnic* with the two microlinguistic measures might have resulted from the fact that it contained more story elements, more heavy verbs, and older adults performed better on comprehension measures for it than for *GDC* (Wright et al., 2011). This underscores the potential clinical utility of CLA, as it may efficiently provide information on multiple components of discourse as analyzed by the LUNA framework (figure 3). Of note, CIUs correlated significantly with core adverb scores for *GDC* only, suggesting that different visual stimuli and parts of speech have different effects on the psychometric properties of CLA, and that these two factors should be considered along with core lexicon scores when interpreting the results.

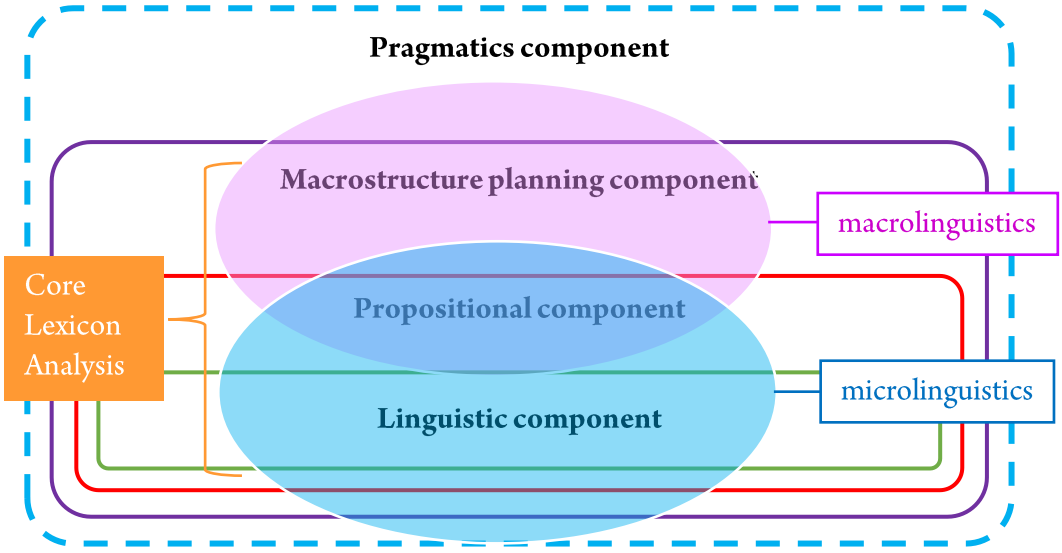

Figure 3 Core Lexicon Analysis (CLA) mapped onto the LUNA framework

On the other hand, there have been no studies on the test-retest reliability of CLA. Test-retest reliability measures the consistency of results over time for an individual when there are no reasons to expect changes in performance. It is an important consideration in developing assessments for clinical and research purposes, because if a measure yields variable results over time, performance on this measure may not accurately reflect the individual's abilities or be sensitive to treatment-related changes. As mentioned above, narrative discourses contain the richest content among the language obtained from discourse elicitation tasks, but narratives are typically longer than the other types of discourse. To obtain the test-retest reliability, one must present the same task for a subject after some specified amount of time, which can pose a significant test burden for some individuals.

The test-retest reliability of several discourse measures has been studied. Brookshire and Nicholas (1994b) evaluated this property for words per minute (WPM), correct information units per minute (CIU/min), and percent correct information units (%CIUs) from discourse samples produced by 20 NBD participants and 20 PWA. The correlation coefficients were 0.90, 0.90, 0.96 for NBD participants, and 0.98, 0.97, 0.98 for PWA. Their study was replicated and expanded by Boyle (2014), who investigated the test-retest reliability of 22 linguistic measures in addition to these using the same stimuli in 12 PWA. Pearson's product–moment correlations (r), standard errors of measurements (SEMs), and minimal detectable change were calculated, and it was found that WPM, CIUs/min, percentage of T-units (defined as a complete idea stated with a verb phrase and at least one noun phrase) with fillers (%TTF), and percentage of T-units with delays (%TD) yielded correlation coefficients >0.90 (r = 0.99, 0.96, 0.90, and 0.94).

Later, Boyle (2015) studied the test-retest reliability of more measures on the word retrieval ability of PWA by following the AphasiaBank discourse protocol

(https://aphasia.talkbank.org/discourse). The measures investigated in the study were phonological errors per minute (PE/min), semantic errors per minute (SE/min), false starts per minute (FS/min), time fillers per minute (TF/min), and repetitions per minute (Rep/min). She found that for the 10 PWA included in the study, the measures with the best test-retest reliability as analyzed by Intraclass Correlation Coefficients (ICCs) were PE/min for complex picture descriptions (ICC = 0.95), SE/min for the Cinderella story retell task (ICC = 0.94), and SE/min for combined picture descriptions-story retell tasks (ICC = 0.96). This work again underscores the importance of specifying both the discourse measures and the type of discourse task when reporting outcomes. Moreover, the correlation coefficients used in Brookshire & Nicholas (1994) was not defined, while Boyle used two different types of correlation coefficients (*r*'s and ICCs) in the two different studies (Boyle, 2014; Boyle, 2015). As such, the values presented in the 3 studies so far were not directly comparable.

Very recently, Stark et al. (2023b) investigated the test-retest reliability of several microlinguistic measures using the AphasiaBank discourse protocol. They found that there were systematic but not statistically significant variations in some measures between test and retest time points (on average 7.79 ± 1.72 days apart) when controlling for group membership and types of discourse tasks. The findings that are relevant to the current project are that 1) measures for fluency/efficiency (CIUs/min, speaking duration, WPM) and for lexical/informativeness (tokens, type/token ratio, %CIUs) were found to have overall superior test-retest reliability compared to measures for syntax (noun/verb ratio and open/closed class ratio) across different discourse tasks for both control and aphasia groups. The ICCs were the highest for %CIUs and CIUs/min across different tasks for both groups, followed by type/token ratio (TTR) and WPM; 2) sample lengths, which was defined as the total number of tokens, showed some correlations with test-retest

reliability. Specifically, for the *Cat Rescue* picture description task, *shorter* samples correlated with higher test-retest reliability of %CIUs for both NBD and aphasia groups. In contrast, for the Cinderella story retell task, *longer* sample length correlated with higher test-retest reliability of CIUs/min, noun/verb ratio, and open/closed class ratio for both NBD and aphasia groups. This presents a more complex picture than an earlier study by Brookshire & Nicholas (1994a), which was on the effect of sample length (as defined by the number of stimuli included) on test-retest reliability. They used a combination of single picture and picture sequence description tasks, discourse prompts about personal information, and procedural description tasks, and analyzed word counts, WPM, CIUs, and %CIUs. It was found that both NBD participants and PWA scored more consistently with greater sample lengths, with higher score variations in the PWA group. Therefore, the relationship between sample lengths and test-retest reliability of discourse measures seem to depend on both how sample length is defined and compared, and the type of discourse elicited.

2.0 Present Study

Recently, we were presented with an opportunity to collect discourse samples in the context of an ongoing study, "Adaptive Trial Scheduling in Naming Treatment for Aphasia" (grant number 1R01DC019325-01A1, funded by the National Institute of Health), on the effects of computerized, adaptive noun picture naming treatment. A key reference for the R01 project was a study conducted by Conroy et al. (2018), where it was found that there was greater improvement in word retrieval at both word and discourse levels when both speed and accuracy of naming were targeted in treatment. Because this encouraging result was obtained using 120 trained and untrained nouns from complex pictures, we sought to conduct the R01 project under matched conditions. Thus, we developed a discourse probe assessment battery that utilized complex pictures and wordless picture books as visual stimuli. Since there are several limitations to the current discourse stimuli, including potential cultural bias in the Cinderella story and artistic styles cater to young children, we selected new stimuli that are more modern, less culturally biased, and more appropriate for adults to elicit discourse.

We reasoned that with the discourse data in hand, we had the opportunity to investigate both the stability of the core noun items and the test-retest reliability of core lexicon scores, specifically that of *core nouns*, over time in the NBD population, for whom the temporal stability of core lexicons should be first established before studying that for PWA. The decision to focus on developing core noun lists and scoring core noun production was twofold: practicality, as this aligns with the direction and work already necessary to complete for the larger study, and clinical utility, as nouns are important targets in current anomia treatments, including Semantic Feature Analysis (Boyle & Coelho, 1995; Lowell, Beeson, & Holland, 1995; Ylvisaker & Szekeres, 1985).

In addition, we measured certain lexical properties of the core nouns generated in this study. Most theories of word retrieval agree that a lexical item has at least two representations in the language processing systems, a conceptual-semantic representation and a phonological representation (e.g., Caramazza, 1997; Dell et al., 1997; Levelt et al., 1999). Quantifying semantic and phonological characteristics in the core nouns may further our understanding of core lexicon analysis; specifically, whether core lexicons are more likely to have certain psycholinguistic properties that have been shown to facilitate single-word picture naming. We selected the following measures:

(1) *Phonological neighborhood density*, the number of lexical items that are phonologically similar in the language (Gahl & Strand, 2016).

(2) *Phonotactic probability*, a combined measure of frequency of individual phonemes and phoneme combinations in a language (Vitevitch & Luce, 1999).

These two measures were selected to investigate the phonological characteristics of core nouns, as previous studies have demonstrated their effects on word retrieval and production. For example, Gordon (2002) and Middleton & Schwartz (2010) found a facilitative effect of higher phonological neighborhood density in the accurate word production of connected speech of PWA. For the NBD population, Vitevitch & Sommers (2003) also identified a facilitative effect of phonological neighborhood density in speech production. In addition, Vitevitch et al. (2004) found that words with higher phonotactic probability were associated with shorter naming latencies. Newman and German (2005) found a significant effect of phonotactic probability on naming accuracy for NBD adolescents and adults. Interestingly, Storkel et al. (2006) studied the effects of each of these variables on word learning in adulthood, and found that higher phonotactic probability correlated with higher numbers of non-words learned; however, error analyses

suggested that phonotactic probability might play a role in acquisition of new words, while phonological neighborhood density might play a role in integrating new words into established lexical knowledge networks.

(3) *Number of features*, the normative counts of semantic features of common lexical items. Specifically, the largest set of existing norms were obtained by asking participants to list features (such as appearance, functions, locations) of selected concepts from semantic memory tasks on a form with 10 blank lines (McRae et al., 2005).

(4) *Contextual diversity*, the count of contexts in which a word appears regardless of its frequency of occurrence (Adelman et al., 2006). Specifically, we used data from an updated model developed by Johns (2021).

These two measures were selected to investigate the semantic characteristics of core nouns. Number of features have been shown to directly correlate with naming latency and accuracy (Lampe et al., 2022; Rabovsky et al., 2016). In terms of contextual diversity, Johns (2021) utilized the discussion posts in the online forum Reddit to derive two updated forms of corpora contextual diversity, discourse contextual diversity (DCD; the count of contexts in which a word appears) and user contextual diversity (UCD; the count of individuals who use a word). Next, he developed computational models using either DCD or UCD, representing different semantic contexts using the Semantic Distinctiveness Model (SD; Jones et al., 2012) and at either the population level (population representation, PR) or the word level (word representation, WR). SD accounts for the number of semantic contexts that a word is used in with weighted values that depends on the semantic similarity of contexts. If a word is used in a similar context as the previous one, the count increases slightly; if it is used in a dissimilar context, the count increases more. PR provides the

count of contexts at the population level, while WR provides the count of contexts for an individual or for a given topic. There were four models tested: DCD-SD-PR, DCD-SD-WR, UCD-SD-PR, and UCD-SD-WR. It was found that UCD-SD-PR provided the best overall fit for explaining lexical strength (such as lexical decision time and naming accuracy) than the other 3 new models. This finding was replicated by Chang et al. (2023), who found that UCD-SD provided better fit than earlier versions of contextual diversity and word frequency. We utilized the UCD-SD-PR model in our study to characterize contextual diversity.

(5) *Confrontational naming difficulty*, a numeric value predicting how challenging it is to retrieve a target word in confrontational naming tasks (Fergadiotis et al., 2019).

Fergadiotis et al. (2019) developed a multiple regression model to estimate the difficulty for a lexical item in confrontational picture naming task that accounted for a substantial amount of variance in item difficulty parameters. They considered the following predictive factors based on a two-stage model of word production, which includes a first stage of accessing the semantic form of a concept and a second stage of activating the phonological form: word length (phonological form retrieval stage), lexical frequency (both stages), and age of acquisition (both stages). A performance variable, the response time for each item on the Philadelphia Naming Test, was added to the model based on empirical utility; its addition led to an increase of the model's predictive power from accounting from 63% (Fergadiotis et al., 2015) to 73% in the 2019 study. Since CLA is proposed to be a measure of word retrieval ability at the discourse level, characterizing the naming difficulty of core lexicons can provide information on whether the NBD population choose lexical items that are more difficult or easy to name in producing discourse. This would be the first step towards studying whether item difficulty at the single word level correlates with that at the discourse level, as measured by CLA.

Finally, as mentioned above, sample length has been found to correlate with test-retest reliability. The direction of such correlation is unclear and may depend on the type of discourse task as well as how sample length is measured. Because this information can guide clinicians in achieving a balance between time efficiency and adequate test-retest reliability during language sample collection, we aimed to describe the strength and direction of the relationship between sample length and test-retest reliability.

In summary, in the context of an ongoing study on computerized adaptive naming treatment, we generated core noun lists for a set of novel discourse stimuli and characterized the lexical properties of core nouns to better understand the information CLA provides about an individual's word retrieval abilities. Given the time efficiency and good concurrent validity of CLA, we aim to investigate its temporal stability, which includes item-by-item stability as calculated by percent agreement and core noun score test-retest reliability as calculated by ICCs and related statistics. Finally, we investigated whether test-retest reliability correlated with sample length. Specific aims and research questions are stated in the next section.

2.1 Specific Aims

1. *Aim 1*: To develop core noun lists for a battery of new visual stimuli, including 15 complex scenes and 11 wordless picture books or chapters, based on aggregated frequency of use (i.e., the counts of the nouns used by participants as a group).
2. *Aim 2*: To evaluate the test-retest reliability of both core nouns and core noun scores from selected visual stimuli by calculating ICCs between two different time points.

3. *Aim 3*: To investigate whether there is a correlation between sample length, as approximated by speaking durations for each type of discourse elicited (picture description and storytelling), and test-retest reliability of core nouns.
4. *Aim 4*: To further characterize lexical properties of the core nouns, including phonological characteristics, semantic characteristics, and predicted naming difficulty at single-word level.

2.2 Research Questions

The proposed research questions related to the above aims are as follows:

1. Does CLA, specifically core noun scores for picture descriptions and story narratives, have adequate test-retest reliability to be utilized clinically? Based on the recommendation from Fitzpatrick et al. (1998), we define adequate clinical utility as ICC > 0.90, which is classified as "excellent reliability" by Koo & Li (2016).
2. Is there a statistical relationship between sample lengths, defined as speaking durations, and the test-retest reliability of CLA?
3. What are the selected lexical properties (phonological neighborhood density, phonotactic probability, number of features, contextual diversity, and confrontational naming difficulty) of the core nouns elicited from selected novel visual stimuli?

3.0 Methods and Procedures

The data for the thesis project was collected in two phases. As mentioned, the data from the first phase is part of that from an ongoing project on computer-based anomia treatment (grant number 1R01DC019325-01A1, funded by the National Institute of Health). It occurred between August 2022 and January 2023. The second phase of data collection was conducted specifically for this thesis, which occurred between March 2023 and April 2023.

3.1 Participants

Thirty-four community-dwelling adults were recruited for the first phase of the study. The recruitment process was approved by the University of Pittsburgh Institutional Review Board and was conducted by (1) posting an electronic flyer on the research group social media page and Pitt+Me®, the university's dedicated site for research participant recruitment, (2) distributing the flyer to aphasia support groups (the rationale was that many care partners of PWA had matching demographic characteristics such as age, ethnicity and linguistic background) and (3) by word-of-mouth to friends and acquaintances. Inclusion criteria included: (1) ages 18 and up; (2) currently community-dwelling; (3) self-perceived as fluent speaker of English; (4) scoring 26 or above on a

cognitive screen, Montreal Cognitive Assessment (MoCA-Basic; Nasreddine et al., 2005)[1]; and (5) have access to a computer and high-speed internet, or can participate in person in Pittsburgh. Exclusion criteria were: (1) current diagnosis of aphasia; (2) prior history of acquired or progressive neurological disease; (3) unmanaged drug/alcohol dependence; and (4) diagnosed severe mood or behavioral disorders.

Participants 3, 4 and 26 did not meet the passing criterion for cognitive screening and were therefore excluded. The remaining 31 participants participated in data collection. However, participants 24 and 33 were lost to follow-up, leading to missing data for some stimuli in the set (please see Appendix B.3 for details of missing data).

There were 24 women (77%) and 7 men (23%) who completed the first phase of the study. The mean age was 54.6 years (range 23–81 years, SD = 16.4 years) and the mean years of education was 16.8 years (range 12–23 years, SD = 3.1 years). The demographic data presented below represents the 31 participants who had completed the norming phase of the study.

[1] In the current study, we followed a cutoff score of 26, which was recommended by Nasreddine et al. (2005). Recent studies have shifted to lower cutoff scores to promote greater racial and ethnic equity (Milani et al., 2018; Ratcliffe et al., 2023).

Table 1 Demographic data for norming participants. The entries for participants who returned for retesting are shaded in blue.

Participant ID	Age	Gender	Race	Ethnicity	Handed-ness	Education in Years (Highest Degree Awarded)
1	67	Male	White	Non-Hispanic/Latino	Right	16 (bachelor's degrees)
2	61	Female	White	Non-Hispanic/Latino	Right	14 (high school diploma)
5	68	Female	White	Non-Hispanic/Latino	Right	18 (master's degree)
6	42	Female	Other	Non-Hispanic/Latino	Right	12 (high school diploma)
7	27	Female	White	Non-Hispanic/Latino	Right	18 (master's degree)
8	65	Male	White	Non-Hispanic/Latino	Right	14 (high school diploma)
9	23	Female	White	Non-Hispanic/Latino	Right	16 (bachelor's degree)
10	35	Female	White	Hispanic/Latino	Right	12 (high school diploma)
11	62	Female	White	Non-Hispanic/Latino	Right	16 (bachelor's degree)
12	73	Female	White	Non-Hispanic/Latino	Right	14 (high school diploma)
13	81	Female	White	Non-Hispanic/Latino	Right	18 (bachelor's degree)
14	43	Female	White	Non-Hispanic/Latino	Right	21 (law degree)
15	62	Female	White	Non-Hispanic/Latino	Right	16 (bachelor's degree)
16	80	Male	White	Non-Hispanic/Latino	Left	22 (master's degree)

Table 1 (Continued)

17	41	Female	White	Non-Hispanic/Latino	Right	16 (bachelor's degree)
18	44	Female	White	Non-Hispanic/Latino	Right	16 (bachelor's degree)
19	72	Female	White	Non-Hispanic/Latino	Right	12 (high school diploma)
20	71	Female	Other	Non-Hispanic/Latino	Right	17.5 (bachelor's degree)
21	62	Female	White	Non-Hispanic/Latino	Right	20 (master's degree)
22	40	Male	African American	Non-Hispanic/Latino	Right	12 (high school diploma)
23	61	Female	Asian	Non-Hispanic/Latino	Ambidextrous	21 (master's degree)
24	29	Female	Asian	Non-Hispanic/Latino	Right	22 (doctorate)
25	46	Male	White	Non-Hispanic/Latino	Left	16 (bachelor's degree)
27	73	Male	White	Non-Hispanic/Latino	Right	21 (doctorate)
28	65	Female	African American	Non-Hispanic/Latino	Right	14 (high school diploma)
29	34	Female	White	Hispanic/Latino	Right	18 (master's degree)
30	65	Male	White	Non-Hispanic/Latino	Right	23 (master's degree)
31	43	Female	Asian	Non-Hispanic/Latino	Right	16 (bachelor's degree)
32	62	Female	White	Non-Hispanic/Latino	Left	16 (bachelor's degree)
33	37	Female	White	Non-Hispanic/Latino	Right	16 (bachelor's degree)
34	59	Female	White	Non-Hispanic/Latino	Right	16 (bachelor's degree)

3.2 Materials

In our study, visual stimuli were selected that 1) contain sufficient amounts of detail to elicit rich language, 2) have been created with artistic style suitable for adult readers, and 3) represent a somewhat complete story.

We recognized that the existing visual stimuli for discourse assessment, including those in the AphasiaBank discourse protocol, *GDC* (Day, 1985), and *Picnic* (McCully, 1984), have been investigated as an ongoing, collaborative effort in standardizing discourse assessment; however, we reasoned that the artistic styles might not be appropriate for adult clients, especially those coming from culturally diverse backgrounds. Therefore, we elected to use new stimuli that have rich details, adequate visual clarity, less cultural bias, and artistic styles that are more appropriate for adolescent and adult viewers. We considered these factors for the following reasons:

1. *Content richness* is a desirable quality. If a visual contains more details, it can provide more opportunities for an individual to use more language.
2. *Visual clarity* is a desirable quality. From a medical perspective, visual impairments may co-occur with aphasia due to brain injury (Smith & Bhutada, 2021). In addition, visual acuity may change during the normal aging process. From a psychological perspective, if there is not enough visual clarity, one can easily feel overwhelmed and frustrated, and therefore perform at a lower level than their true ability.
3. *Infantilization* is an undesirable quality. Even though in this thesis project, the participants were not living with aphasia, we searched for materials with the ultimate goal of using them in aphasia assessment. Adults with aphasia often encounter infantilizing treatments from others, which stem from either the misunderstanding about aphasia, that their language impairments reflect their intellectual impairments, or negative attitudes

towards disabilities (Howe et al., 2008). Considering that this is a barrier to social participation for PWA, we seek to prevent a perception of being infantilized or patronized, so that PWA and their clinicians can build positive, therapeutic relationships.

4. *Cultural bias* is an undesirable quality. A culturally specific story or reference does not accurately reflect a person's language ability because background knowledge and familiarity can have confounding effect on the person's narrative performance. We seek to minimize cultural bias so that PWA from diverse backgrounds can benefit from our findings.

To select items meeting the above criteria, research assistants first conducted broad internet search of complex scenes and wordless picture books and collected images for team discussion. Next, potential pictures and books were determined via discussion before purchasing the physical copies. After obtaining the books, research assistants scanned all pages and a second round of team discussion was conducted to select suitable pages from the books. Finally, the selected pages were rated by all research team members considering all the above criteria, and those with best ratings were selected as discourse visual stimuli.

The final list of visual stimuli included 15 complex scenes and 11 self-contained stories from wordless picture books.[2] Since a relevant study by Conroy et al. (2018) utilized complex scenes from "*Where's Waldo?*" book series as visual stimuli to compare generalization of naming treatment efficacy with and without cues and deadlines, pages from the same book series were selected as 11 out of 15 the final complex scenes. All but one of the wordless picture books were

2 The self-contained stories were either selected pages of each chapter of the book, *the Arrival*, or the rest of the listed books in their entirety.

presented in their entireties. The exception was the book *the Arrival* (Tan, 2007): due to its length, only selected pages from each chapter were included. This list is included in Appendix B.

3.3 Procedure

The procedure was conducted in a HIPPA-compliant zoom environment for all participants. Prior to data collection, details about the project including aims, risks and benefits were explained, and informed consents from participants were obtained via the secure online platform REDCap (Harris et al., 2009). Next, the MoCA-Basic (Nasreddine et al., 2005) was administered with minor modifications for telehealth presentation (Loring et al., 2023)[3] to screen for the confounding effects of potential cognitive impairments. Visual stimuli were presented as slides virtually in a randomized order unique to each participant. Participants received the following prompts:

> For complex scenes: "*Imagine you are describing the scene to someone who cannot see the picture. Describe what you see and what is happening in as much detail as possible, and please try to take between 3 and 5 minutes to do so*."
>
> For wordless picture books: "*Imagine you are telling the story to someone who cannot see the pictures. Describe what you see and what is happening as you go through the story. Each story is around 30 pages long. In terms of the level of detail, please try to take roughly 10 to 15 minutes to tell the story*."

[3] Our modifications were consistent with those outlined in Loring et al. (2023) with one minor difference: we presented all figures for naming on the same slide, while Loring et al. presented the figures one by one on separate slides.

Because of examiner error, participant 5 and participant 7 received additional visual stimuli in the form of 1 and 2 unabridged wordless picture book chapters respectively,[4] whereas all other participants viewed abridged chapters. The relevant parts of the language samples from participants 5 and 7, namely when they talked about the pages that were shown to all participants, were included in the analyses. Finally, participants who were lost to follow-up completed variable numbers of items in the set (26 items in total): participant 24 completed 5 items and participant 33 completed 14 items.

After a mean interval of 149 days (range 86 to 162 days), 15 out of 31 participants returned for a second iteration of the data collection procedure (the entries for participants who returned for the retest are shaded in blue in Table 1). We intended to present the items in the same order for each participant to keep the variable of presenting order constant; however, participant 1 and participant 10 received items in new randomized orders, and participant 1 was also shown a minimally different order (two items were swapped) due to examiner errors.[5]

3.4 Data Preparation and Analysis

All videos and audios were recorded via Zoom. The recordings were fully and automatically transcribed by Microsoft Stream in orthography. Three research assistants including the author reviewed the transcripts and corrected mis-transcribed words. In addition, the research

4 Participant 5 received the entire chapter 2 of *the Arrival* (Tan, 2007), and participant 7 received the entire chapters 3 and 6 of the same book. All other participants received selected pages of these chapters.

5 The affected items are two abridged chapters of *the Arrival* (Tan, 2007), "Arrival 2" and "Arrival 4".

assistants marked the start and end of the transcripts relevant to each visual stimulus to be analyzed (coded as "#START_[stimulus name]" and "#END_[stimulus name]"), while marking the irrelevant information to be excluded from analysis using "{ }" brackets. The irrelevant information consisted of conversations on using zoom, unexpected interruptions in the sessions, and requests for turning the page of a wordless picture book. The manually corrected transcripts were then analyzed in R (Version 4.3.2) using scripts written in RStudio 2023.12.1+402 "Ocean Storm" Release (4da58325ffcff29d157d9264087d4b1ab27f7204, 2024-01-28) for Windows. All scripts written in R can be found at our OSF project site .[6]

From each participant's language samples, the relevant segments were extracted, converted to lower-case letters, and then tokenized. Tokenization is the process of segmenting a text into smaller parts, or "tokens". The *tokenizers* package was utilized (Mullen et al., 2018), where the user can select the type of tokens for data output. The type, *regular expression* (regex), was selected, which allows the user to determine the rule for segmenting. In this case, we set the rule as segmenting by white spaces, which would return both regular and contracted words as tokens. To prevent compound words from being coded as separate single words, the space between parts was deleted; for example, "ice cream" was replaced with "icecream". All punctuation marks were then excluded. Finally, the tokens were grouped under each stimulus.

[6] Please use this link to access OSF site: https://osf.io/v7asz/?view_only=099fc05e592047bb81be7040e31c64a9. It will be published once reviewed and revised.

3.4.1 Core Noun List Generation

There are currently two approaches for developing core lexicon lists. One is to select lexicons used by a set *percentage* of NBD participants. This approach was used by Dalton & Richardson (2015) and Alyahya et al. (2022), whose cutoff percentages were 50% and 75% respectively. The other approach is to include a set *frequency* of use, which refers to the numbers of appearances in the language samples when considering NBD participants as one or multiple groups. This approach was used by MacWhinney et al. (2010) and Kim et al. (2019), who included the 10 and 25 most frequently used lexical items respectively. To evaluate which criterion better measured the underlying construct, namely the word retrieval ability at the discourse level, Kim et al. (2022) analyzed language samples elicited from wordless picture book story narrative tasks using confirmatory factor analysis. Based on their model, both approaches of determining core lexicons measured the construct to similar extents. However, the authors found that measurement errors of content words, including nouns, verbs, adjectives, and adverbs, were computed to be higher using the percentage criterion. They noted that the difference in measurement errors was not significant based on their computational model. In addition, because measurement errors can come from multiple sources, factors such as group demographics and levels of education should be considered alongside the criterion for core lexicon inclusion.

To minimize measurement errors, we adopted the criterion based on *frequency* in determining core nouns. This was accomplished as follows. The output from tokenization and grouping for all 31 participants were tabulating using two columns: stimuli and response. The combined responses for all participants for each stimulus occupied one row. The nouns in the response for each stimulus were tagged using *TreeTagger* (Schmid, 1997) package in R and the unique lemma were counted. For each stimulus, the lexical items were ranked from most

frequently used to least frequently used. Next, a separate table was created where each stimulus was associated with the desired number of core nouns based on their type; in other words, a complex scene was associated with the number "10" and a wordless picture book section was associated with the number "25". These numbers were selected because they have been used by other authors; however, there are not yet published studies investigating the optimal numbers of core lexicons for different discourse tasks. This table of stimuli served as a key for generating core noun lists, yielding the top 10 most frequently used nouns for each complex scene and the top 25 most frequently used nouns for each wordless picture book section.

3.4.2 Core Noun Scoring

The core noun scores were determined for the 15 participants who completed both testing and retesting using the core noun checklists generated from the language samples of all 31 participants. In other words, core noun checklists were generated based on the language samples of all participants, but only the test-retest participants were scored. To score core noun use, the nouns in the hand-corrected discourse samples from each participant were tagged using *TreeTagger* (Schmid, 1997) package and grouped by stimuli in a separate table. Next, core nouns were compared against all the nouns used by a participant for each stimulus, and core noun use was scored according to the established scoring rules listed below (Kim & Wright, 2020a):

1) A participant earned 1 point for each core lexicon that they used regardless of how many times they had used the lexicon in their discourse.
2) There was no point awarded for synonyms.
3) Inflected forms (e.g., plurals) of a core lexicon were credited.
4) Raw scores were converted into percentage using this equation:

$$\text{score} = \frac{\text{nouns on the list used by the participant}}{\text{total number of nouns on the list}} \times 100$$

Whether synonyms of the core nouns should be credited can be rightly a point of contention. When comparing the Cinderella stories told by PWA with those by NBD participants, MacWhinney et al. (2010) noted that the nouns produced by PWA were less specific to the story. For example, PWA were more likely to use "shoe", "girl", and "man", while NBD participants were more likely to use "daughter/stepdaughter", "fairy", and "godmother". For this reason, there is concern that scoring synonyms, especially those less relevant to a given discourse task, may undermine the sensitivity of CLA in distinguishing PWA from the NBD population. The only exceptions reported in the studies were for the words "father" and "mother", for which all variations (such as "dad", "daddy", "mom", and "mommy") were given credits (Dalton et al., 2020). When both synonyms meet the core lexicon criterion, both are included on the checklists, so an individual can still receive credits for using either one of the synonyms.

It is beyond the scope of this thesis to delineate the complexity of synonyms — not all synonyms have the same level of semantic richness and topic relevance, and it would require in-depth analysis to develop a psycholinguistically robust framework to include one synonym but not the other (e.g., including a separate procedure to determine how relevant a synonym is to a discourse topic). We elected to not award points for all synonyms to be consistent with previous studies, except for variations of "father" and "mother".

3.4.3 Test-Retest Reliability

Thereafter, the ICCs and the 95% confidence intervals (CIs) between the core lexicon scores at time point 1 and that at time point 2 were calculated using the *irr* package in R (Gamer

et al., 2012). According to the guidelines for selecting and reporting test-retest reliability (Koo & Li, 2016; Liljequist et al., 2019), we calculated the ICC estimates and their 95% CIs based on a single-measurement/rater, absolute-agreement, two-way mixed-effects model (McGraw & Wong, 1996). This form was denoted as ICC (3,1) by Shrout and Fleiss (1979).

It is known that ICCs are larger for more heterogenous samples, whereas standard errors of measurement (SEMs) are less affected by group heterogeneity. Therefore, SEM was calculated to complement the analysis for reliability using this equation: $\text{SEM} = \text{standard deviation} \times \sqrt{1 - \text{ICC}}$, where the standard deviation was a pooled value of the standard deviations in core noun scores at both time points: $\text{standard deviation} = \sqrt{\frac{\text{SD}_1^2 + \text{SD}_2^2}{2}}$ (De Vat et al., 2011).

Using SEM, the minimum detectable change (MDC) for 95% CI was calculated as $\text{MDC}_{95} = \text{SEM} \times \sqrt{2} \times \text{z}$, where z = 1.96 (Stratford, 2004). An MDC at 95% CI provides information on the minimal amount of change that must be detected to conclude that the change is not due to measurement error 95% of the time.

3.4.4 Percent Agreements

Percent agreement values were calculated using this equation: $\text{percent agreement} = \frac{\text{number of individual core nouns used in both time points}}{\text{number of individual core nouns used in time point 1}}$, where individual core nouns were determined based on the numbers of appearances of nouns in a participant's language sample (i.e., the same procedure as the one used to determine core nouns for the group). The nouns used at both time points were identified for each participant.

3.4.5 Sample Lengths

Sample lengths were approximated by the speaking durations. An R script was developed to extract the start and end timestamps for the part of an individual's language sample for each discourse item and calculate the discourse duration. The Shapiro-Wilk test was applied to sample lengths by stimuli, revealing that the values are not normally distributed. Therefore, the medians of sample lengths were used in the subsequent calculation for correlation. Further Shapiro-Wilk tests indicated that the medians of sample lengths were not normally distributed, though ICCs were. Because the medians of the sample lengths were not ranked or normally distributed, Spearman's ρ instead of Pearson's product-moment correlation coefficient was calculated to identify whether there was a correlation between the medians of sample lengths and ICCs. Similar analysis was performed for each participant, where the Shapiro-Wilk test was conducted to assess the normality of the distributions of both sample lengths and percent agreements, revealing that none of the sample lengths for each participant was normally distributed. Thus, Spearman's ρ was calculated for each participant.

3.4.6 Lexical Properties

Phonological neighborhood density and phonotactic probability values were obtained from the Irvine Phonotactic Oline Dictionary (IPhOD version 2.0, available at https://www.iphod.com/; Vaden et al., 2009). Phonotactic probability takes into consideration both biphoneme frequency, the frequency at which a two-phoneme pair occurs in the language, and positional probability, the probability that a phoneme appears at a specific position of a word. The specific values of phonological neighborhood density and phonotactic probability were calculated based on

unstressed syllables, meaning that vowels were not distinguished by whether they carried main stress prosodically. In addition, the values of positional probability were compared among words of the same length (i.e., length-constrained) and weighted with context count measures in $SUBTLEX_{US}$, which was another measure for word frequency in $SUBTLEX_{US}$ that minimized the confounding effect on proper nouns (Brysbaert & New, 2009). These forms were chosen to minimize bias due to dialectal differences in prosody and minimize the impact of word length and word frequency on phonotactic probability; the IPhOD abbreviations of them were unsDENS (for phonological neighborhood density), unsCBPAV (for biphoneme probability), and unsCLCPOSPAV (for positional probability).

The numbers of semantic features were obtained from the norms established by McRae et al. (2005), and contextual diversity was represented by the UCD-SD-PR values from the data collected by Johns (2021). The confrontational naming difficulty was calculated using the model developed by Fergadiotis et al. (2019): $-2.012 - 0.391 \times \log_{10}\text{CD} + 0.15 \times \text{AoA} + 0.136 \times (\text{number of phonemes}) + 1.482 \times (\text{response latency})$, where $\log_{10}$CD values were obtained from $SUBTLEX_{US}$ corpus (Brysbaert & New, 2009), age of acquisition (AoA) values from Kuperman et al. (2012), response latencies from Székely et al. (2003), and numbers of phonemes from IPhOD version 2.0. The numbers of phonemes not included from the downloadable IPhOD data and the response latencies missing in Székely et al. (2003) were obtained from the English Lexicon Project (ELP; Balota et al., 2007; available at https://elexicon.wustl.edu/). It is worth noting that unlike the naming tasks in Székely et al. (2003) and Fergadiotis et al. (2019), which utilized pictures as stimuli, the ELP naming tasks displayed the word forms for the participants to name. The change in stimuli type might have had an impact on naming latencies, and therefore the naming difficulty calculated. When naming latencies were not available from Székely et al. (2003) or the ELP, item

difficulty was calculated based on a previous study by Fergadiotis et al. (2015), which was shown to explain 63% of the variance in PNT naming performance by using the following equation: $-1.22 + 0.15 \times (\text{number of phonemes}) - 0.36 \times \log_{10}\text{CD} + 0.21 \times \text{AoA}$. Finally, for the two compound words in the core noun lists, “ice cream” and “palm tree”, values for each constituent word were included when that for the compound word was not available. Item difficulty values for compound words were not calculated due to the lack of predictive model for them.

4.0 Results and Discussion

This study primarily aims to evaluate the test-retest reliability of CLA, a discourse measure that requires little transcription and therefore can be more readily included in clinical assessment of language skills of PWA. Test-retest reliability is one aspect of reliability, and reliability can be defined as "the degree to which the measurement is free from measurement error" (Mokkink et al., 2010, p.743). If a measurement is not adequately reliable, then the changes shown on the measurement will not necessarily reflect changes in the skills or knowledge being measured. To evaluate test-retest reliability, we calculated ICCs to compare them with existing benchmarks. We also calculated individual percent agreement and extracted the nouns that an individual used at both time points. Our findings regarding the test-retest reliability of CLA are addressed in section 4.1.

The secondary aims include to determine whether there is a correlation between sample lengths and test-retest reliability and to describe the phonological, semantic, and behavioral (specifically with regards to confrontational naming) characteristics of core nouns. Clinicians are often challenged with deciding when a language sample is long to ensure that it provides adequate information, while remains short enough that it can be analyzed within a reasonable timeframe. Therefore, understanding whether sample lengths may affect the test-retest reliability of a measure can guide such clinical decision-making. This aim is addressed in section 4.2. Finally, the lexical characteristics of core nouns are included in section 4.3 as a step towards better understanding of the underlying construct of CLA; in other words, whether CLA measures an individual's ability to access lexical items with selected psycholinguistic characteristics.

4.1 RQ1: Does CLA, specifically core noun scores for picture descriptions and story narratives, have adequate test-retest reliability to be utilized clinically?

The ICCs for these data ranged from –0.517 (95% CI [–0.875, 0.033]) to 0.695 (95% CI [0.292, 0.887]), which can be classified as "poor to moderate" according to Koo & Li (2016). As such, the test-retest reliability of core noun scores for this set of stimuli did not meet the cutoff value (0.90) for concluding that an instrument is reliable to assess individuals (Fitzpartick et al., 1998); in fact, the highest ICC, 0.695, was just below the criterion (ICC ≥ 0.70) for concluding that an instrument is reliable to assess a group of individuals.

ICCs are influenced by factors other than test-retest reliability. With a small sample size of 15, the results may reflect the lack of heterogeneity in the group. Perhaps the more informative values are MDC_{95}, which ranged from 19.3 to 67.2 with a median value of 30.25, indicating that a minimal change of 19.3 in core noun scores must be observed to conclude that a true change in core noun retrieval ability has taken place.

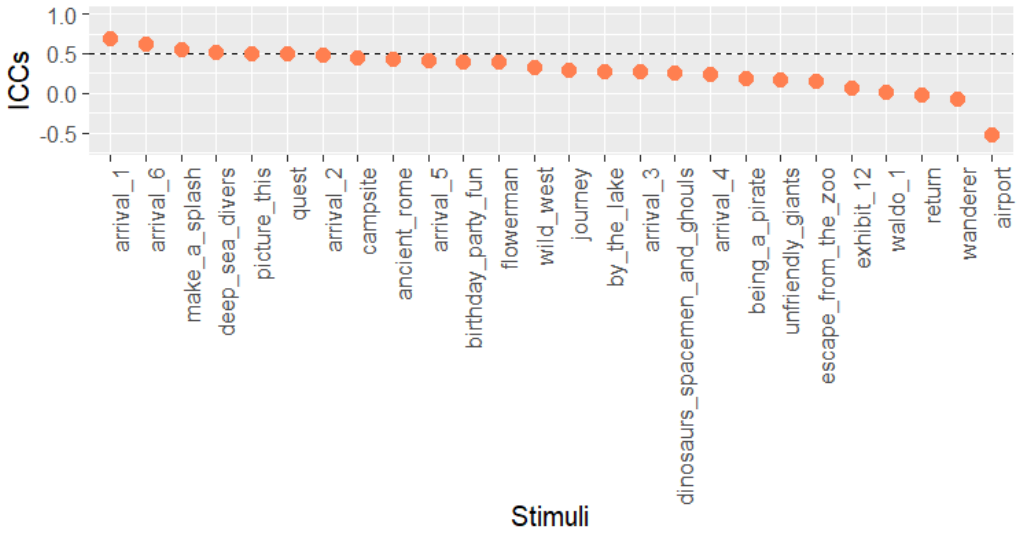

Figure 4 ICCs plotted by stimuli . A line at ICC = 0.5 (below which the values are classified as "poor") is shown for reference.

Table 2 ICCs, SEMs, and MDC_{95} for all stimuli

Stimuli	ICCs	95% CIs	Classifications	SEMs	MDC_{95}
Airport	-0.517	[-0.875, 0.0333]	poor	7.19	19.9
Ancient Rome	0.427	[-0.0299, 0.753]	poor to good	7.76	21.5
Arrival 1	0.695	[0.292, 0.887]	poor to good	14.1	39.1
Arrival 2	0.491	[-0.0324, 0.797]	poor to good	18.7	51.8
Arrival 3	0.268	[-0.203, 0.662]	poor to moderate	9.65	26.8
Arrival 4	0.24	[-0.269, 0.654]	poor to moderate	24.2	67.2
Arrival 5	0.414	[-0.125, 0.759]	poor to good	7.16	19.8
Arrival 6	0.624	[0.176, 0.856]	poor to good	14.4	40
Being a Pirate	0.188	[-0.367, 0.633]	poor to moderate	14.8	41
Birthday Party Fun	0.398	[-0.145, 0.751]	poor to good	7.48	20.7
By the Lake	0.276	[-0.142, 0.654]	poor to moderate	10.6	29.5
Campsite	0.452	[-0.0197, 0.77]	poor to good	6.95	19.3
Deep Sea Divers	0.512	[0.0182, 0.805]	poor to good	9.5	26.3
Dinosaurs, Spacemen, and Ghouls	0.265	[-0.307, 0.681]	poor to moderate	14	38.8
Escape from the Zoo	0.148	[-0.419, 0.611]	poor to moderate	11.2	31
Exhibit 12	0.0593	[-0.5, 0.554]	poor to moderate	15	41.6
Flowerman	0.397	[-0.155, 0.751]	poor to good	8.98	24.9
Journey	0.285	[-0.283, 0.692]	poor to moderate	9.22	25.6
Make a Splash	0.55	[0.106, 0.818]	poor to good	15.6	43.4
Picture This	0.506	[-0.0105, 0.804]	poor to good	8.68	24
Quest	0.493	[-0.0147, 0.798]	poor to good	11.9	32.9
Return	-0.0247	[-0.571, 0.497]	poor	13	36
Unfriendly Giants	0.172	[-0.396, 0.625]	poor to moderate	7.41	20.5
Waldo 1	0.0131	[-0.509, 0.514]	poor to moderate	8.02	22.2
Wanderer	-0.0662	[-0.564, 0.453]	poor	15.6	43.2
Wild West	0.327	[-0.232, 0.714]	poor to moderate	12	33.2

An important consideration for reliability testing is the stability of the underlying construct of a measurement. If the function or skill being measured changes over time, the measurement outcomes will change accordingly. Kim et al. (2019) calculated percent agreements of core lexicons between different NBD cohorts grouped by age, and found that for core nouns, the percent agreements ranged from 56% between those in their 50s and those in their 80s, to 80% between several pairs of cohorts. This suggests that lexical selection in discourse, which is the underlying construct driving CLA, changes over the lifespan. However, the rate at which this ability changes is unclear. In our study, the time elapsed between testing and retesting was 3–5 months, during which changes in lexical selection might have occurred. Though the magnitude of such changes is underexplored, Kim et al. (2022) found that the core lexicon scores for PWA significantly improved from acute to chronic phases of aphasia: over a period of 4 to 14 months, the mean scores increased from 38.5 to 61.9.

At the individual level, the percent agreement values of core nouns were computed. As shown in figure 5, the values ranged from 0.0 to 0.8, with most median values at 0.5 and 0.6. These values suggest that each individual may change how often they use nouns in a discourse task: the identities of the most frequently used nouns depend on both the individual and the task.

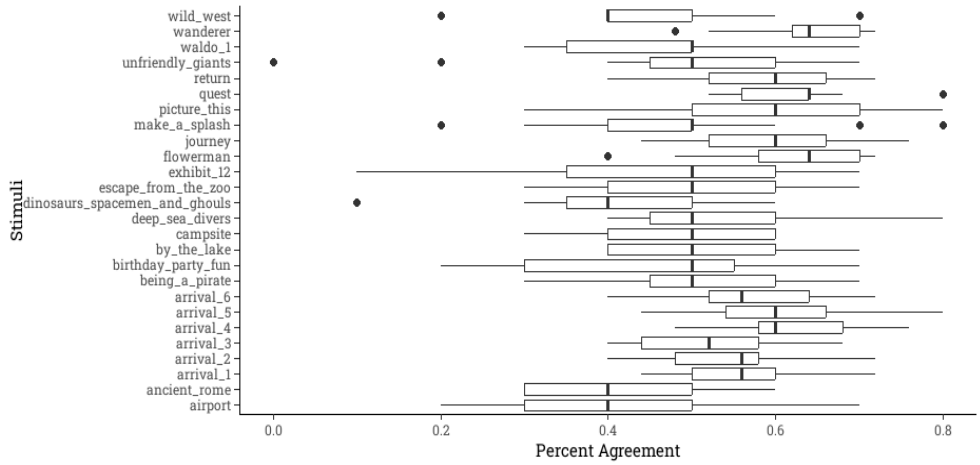

Figure 5 The distribution of percent agreement plotted by stimuli

4.2 RQ 2: Is there a statistical relationship between sample lengths and the test-retest reliability of CLA?

At a group level, Spearman's ρ between sample lengths, approximated by speaking durations, and ICCs was calculated to be –0.0885 with a *p*-value of 0.666 (figure 6). This suggests that there is no significant correlation between sample lengths and the test-retest reliability of core noun scores. For individuals, the values of the Spearman's ρ ranged from 0.14 for participant 19 to 0.72 for participant 30 with *p*-values < 0.05 in all but the 3 cases where Spearman's ρ values were the smallest (figure 7; please see Appendix C for tabulated data). The small to moderate positive Spearman's ρ values suggest that at the individual level, longer samples tend to have higher test-retest reliability. This seemingly inconsistent result can be explained by (1) the individual variations in discourse production, i.e., each participant's discourse style, and (2) the different dependent variables considered, namely ICCs vs. percent agreements.

Some of the common definitions of sample lengths are as follows: total number of tokens (e.g., Stark et al., 2023b), total number of words (e.g. Spencer et al., 2023), total number of C-units (defined as an independent clause with all of its subordinate clauses; e.g., Greenhalgh & Strong, 2001), number of utterances (defined as an expression of thought indicated by intonation, a pause, an interruption, or the start of a different thought; e.g., Pavelko et al., 2020) and speaking durations (e.g., Boyle, 2015). We selected to approximate sample lengths using speaking durations because measuring time is more feasible to perform during an evaluation or treatment session. An alternative definition of sample lengths may be more suitable when eliciting language samples from PWA, especially those who present with slow, effortful speech.

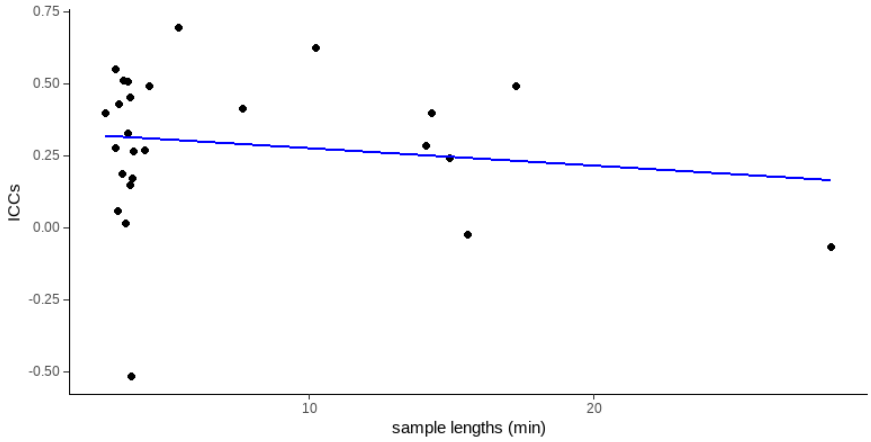

Figure 6 ICC plotted against sample lengths at group level . Each data point represents the ICC for each participant.

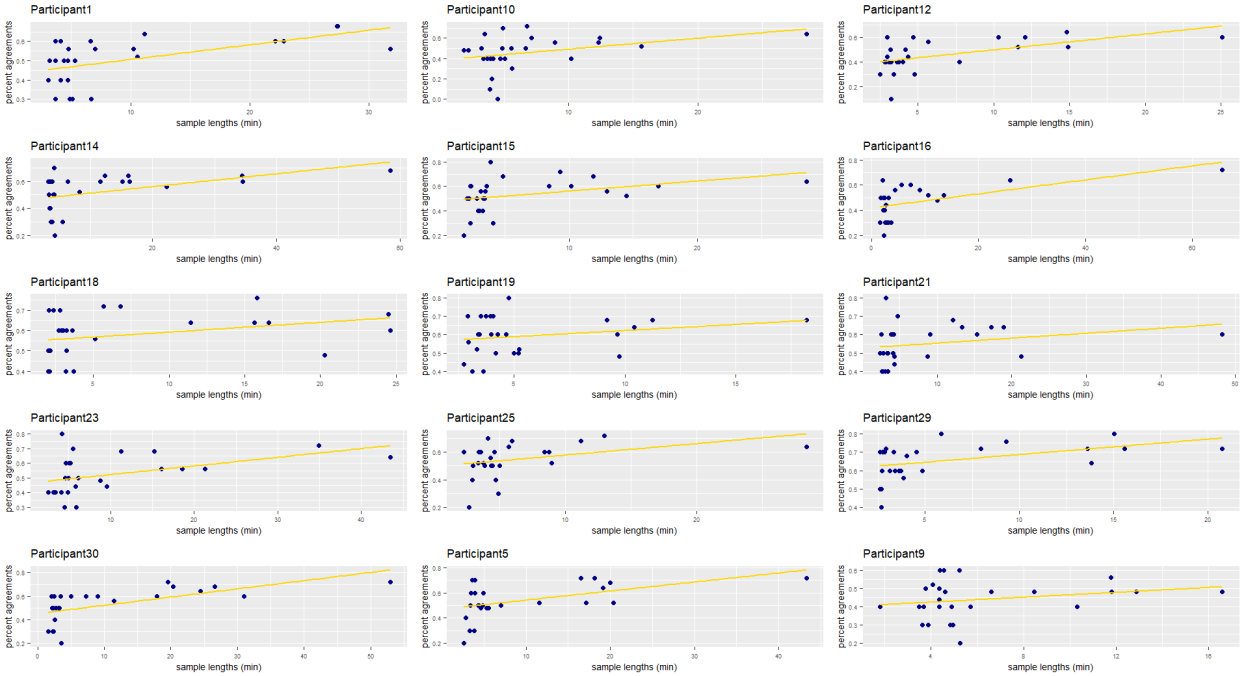

Figure 7 Percent agreements plotted agains sample lengths for each participant . Each data point represents the percent agreement of core nouns used for each stimulus.

4.3 RQ3: What are the selected lexical properties of the core nouns elicited from selected novel visual stimuli?

The summary of lexical properties of core nouns grouped by stimuli are listed below in table 3. For comparison, the values of phonological neighborhood density, biphoneme probability, positional probability, and number of features from selected references (Middleton & Schwartz, 2010; Rabovsky et al., 2016; Storkel et al., 2006; Vitevitch et al., 2004) are also included. These references were selected, because significant differences were reported between the stimuli with high numeric values of lexical characteristics and those with low numeric values of lexical characteristics. It is worth noting that there were different criteria for classifying the values as

"high" or "low" depending on the study, and that the forms of phonotactic probability used were not specified (e.g., calculated using unstressed or stressed syllables). In comparison, the current set of stimuli yielded core nouns with moderate to high phonological neighborhood densities, low to moderate biphoneme probabilities, low positional probabilities, and low to moderate numbers of semantic features. Given the known positive correlations between these lexical properties and naming latencies and naming accuracy, this may suggest that the core nouns obtained in our study were more difficult to name.

Table 3 Reference Values for Lexical Properties

	Reference	High (mean)	Low (mean)	Average Values from Current Set of Stimuli
Phonological Neighborhood Density	Storkel et al., 2006	14	5	18.25
	Vitevitch et al., 2004	20.7	17.6	
	Middleton & Schwartz, 2010	23.04	10.12	
Biphoneme Probability	Storkel et al., 2006	0.0056 – 0.0066	0.0010 – 0.0017	0.00353
	Vitevitch et al., 2004	0.013	0.005	
Positional Probability	Storkel et al., 2006	0.16 – 0.15	0.009	0.0670
	Vitevitch et al., 2004	0.206	0.153	
Number of Features	Rabovsky et al., 2016	21	8 – 10	13.98

The lexical properties for the core noun lists of each visual stimulus, along with the core noun lists from *GDC* and *Picnic* normed from individuals in their 50s and 60s (Kim et al., 2019), are included in table 4. As shown below, the calculated naming difficulties of core nouns from the current stimuli were higher than those from *GDC* and *Picnic*, suggesting that the core nouns in this

study were more difficult to name than those in *GDC* and *Picnic*. Detailed lexical properties can be found in Appendix D.

Table 4 Summary of lexical properties by stimuli

Stimuli	Average Num. of Phonemes	Average Phonological Neighborhood Density	Average Biphoneme Probability	Average Positional Probability	Average Naming Difficulty	Average Contextual Diversity	Average Num. of Features
GDC (50s)	3.87	17.84	0.002669	0.061343	-1.131968	3557.297	17.25
GDC (60s)	3.96	18.72	0.002791	0.059595	-1.121911	3567.73	17
Picnic (50s)	3.9	19.6	0.002544	0.062764	-1.294319	3604.842	12.6
Picnic (60s)	4.14	16.8	0.002365	0.060364	-1.254235	3565.074	12.6
Airport	4.4	11.8	0.0031	0.06245	-0.74957	3048.111	14.33
Ancient Rome	4.5	19.1	0.005566	0.064975	-0.65284	3478.049	17.67
Arrival 1	3.92	22.46	0.002625	0.062913	-0.85477	4004.983	14.56
Arrival 2	4	19.88	0.003219	0.067209	-0.78865	3955.43	13
Arrival 3	4	21.56	0.003153	0.068057	-0.9704	3994.758	14.5
Arrival 4	3.88	19.32	0.003431	0.068285	-0.86656	4111.978	13.25
Arrival 5	4	21.96	0.003379	0.074012	-0.83468	4205.719	12.71
Arrival 6	4.28	20.67	0.003074	0.06283	-0.75475	3777.197	14.4
Being a Pirate	3.89	16.88	0.003877	0.073005	-0.53408	2893.289	13.33
Birthday Party Fun	4	16.2	0.002881	0.06298	-1.24404	4513.434	14
By the Lake	3.6	26	0.002611	0.070521	-0.97072	3377.522	13.5
Campsite	3.89	18.44	0.003903	0.065785	-0.97435	3699.408	12.33
Deep Sea Divers	3.6	22.1	0.00236	0.07005	-0.86946	3782.443	11
Dinosaurs, Spacemen, and Ghouls	4.67	14.13	0.003199	0.063988	-0.64435	3457.276	–
Escape from the Zoo	5.11	9.44	0.004108	0.057582	-0.26146	1914.08	17.33
Exhibit 12	4.4	15.4	0.003992	0.074293	-0.69559	4217.594	11.5
Flowerman	3.8	16.64	0.00395	0.065404	-1.07829	4681.603	15.8
Journey	3.72	19.64	0.003367	0.066257	-0.9824	4007.742	14
Make a Splash	3.9	16.9	0.002707	0.05782	-1.02045	3749.433	13
Picture This	4.1	23.33	0.004183	0.083246	-1.07487	4502.536	–
Quest	4.08	17.68	0.003502	0.06662	-0.69298	3910.216	13
Return	3.8	23.13	0.00283	0.067021	-0.86757	3994.596	12.5
Unfriendly Giants	4.1	13.5	0.005142	0.067097	-0.89213	3859.523	16
Waldo 1	4.4	15.5	0.00509	0.066398	-0.84006	4106.501	15.5

Table 4 (Continued)

Wanderer	3.84	23.95	0.003539	0.071664	-0.88114	3893.158	12
Wild West	4.4	8.8	0.003045	0.060538	-0.7193	3605.416	16.33

5.0 Conclusions

Under our experimental conditions, the test-retest reliability of core noun scores as measured by ICCs was poor and core nouns for the selected stimuli varied considerably over time for each individual. There was a small to moderate positive correlation between sample lengths and ICCs for most participants as measured by Spearman's ρ, suggesting that longer discourse tasks may provide more reliable information on the word retrieval ability at discourse level.

5.1 Limitations

There are several limitations to this work. First, even though we have strived to improve the quality of visual stimuli, the stimuli used in this study still do not include protagonists and plots from culturally and ethnically diverse backgrounds. The best wordless picture books featuring minoritized peoples seem to be more appropriate for young children. We recognize the importance of representation and therefore we hope to conduct future study with more culturally inclusive materials. A future study can focus on the perception of PWA towards discourse assessment stimuli, so that clinicians and researchers can study stimuli with the characteristics that PWA value.

Secondly, the time elapsed between testing and retesting was longer than similar studies on test-retest reliability (for comparison, the time elapsed was 7–10 days in Brookshire & Nicholas' 1994 study, and 2–7 days in both of Boyle's 2014 and 2015 studies, and within 2 weeks in Stark et al.'s 2023b study). Since the rate of change in the word retrieval ability for discourse

production, the underlying construct of CLA, is understudied in NBD populations, it was unclear how much it has changed in our participants over the span of months. Moreover, we included a small sample size (N = 15) in our study. Since a smaller group is less likely to be heterogeneous, ICC estimates were likely larger.

In addition, the demographic characteristics of our participants were not well matched with those of PWA, especially in terms of gender and ethnicities. Our participants were predominantly white and female, and therefore our findings may not be applicable to diverse populations.

In terms of procedure, the language transcripts were automatically generated using speech-to-text software developed by Microsoft. Even though the transcripts were thoroughly reviewed while re-watching the video and all nouns were hand corrected by trained research assistants, the accuracy of these transcripts may not be suitable for other linguistic analyses.

There were two major differences in our procedure compared to existing protocols: (1) participants were not given opportunities to read the wordless picture books before telling stories about them, which might have affected lexical decisions such as using “daughter” or “girl”. (2) There were time constraints included in the prompts, directly affecting our measurement of speaking durations. We recognized that these differences had led to decreased external validity of our findings; in future studies, we will follow a similar discourse protocol as those in existing studies, allowing material review and leaving out time constraints.

Finally, this work only focuses on core nouns given the focus of the overall study and the quality of available data. As mentioned above, Kim & Wright (2020b) found that core lexicons had different concurrent validity values with certain macrolinguistic measures depending on their parts of speech of speech. As such, the findings cannot be generalized to all core lexicons.

5.2 Future Directions

The current study is best viewed as a small step towards developing a set of valid, reliable, and functional measures for discourse analysis (i.e., COS for discourse) in aphasiology. There is much work to be done to realize this long-term goal. Future work can be described in terms of proximal and distal research questions: proximal questions are direct follow-ups of current study, while distal questions are more general conceptual inquiries. Based on the limitations of this study, we propose the following 2 proximal research questions for future investigation:

1) *How does the test-retest reliability of core lexicons from other parts of speech compare with that of core nouns*?

The core lexicon checklists for other parts of speech, such as function words and verbs, can also be generated from the study data. Kim and Wright (2020b) demonstrated that CLA of different parts of speech correlated variably with other discourse measures. This difference in concurrent validity suggests that other psychometric properties may depend on the parts of speech of the core lexicons, even when the stimuli and tasks are kept constant.

2) *How does the core noun test-retest reliability change under different experimental conditions*?

As discussed in the previous sections, the present study included a small sample size and a long interval between testing and retesting. It is likely that core noun test-retest reliability as measured by ICCs will change with larger sample sizes and shorter test-retest time intervals. Moreover, studying larger samples can increase external validity, and shortening test-retest time intervals, especially to 2–14 days, can facilitate more direct comparison with the data from other reliability studies.

Two examples of distal research questions are as follows:

1) *What is the relationship between making sense of external materials/experience and word retrieval ability at discourse level*?

The participants in our study did not have the opportunity to view the wordless picture books in their entirety before producing discourse, and therefore had limited understanding of the stories before telling them. In contrast, in everyday life, usually we have already known the sequence of the true or fictional events and speculated the relationships between them before we tell stories about them. Studying how the subjective understanding of a situation or a series of events affects lexical decisions for discourse can inform how a discourse elicitation task should be presented to minimize or account for the effects of perceptual, cognitive, and pragmatic deficits, as well as whether and how much emphasis should be placed on comprehension strategies in aphasia treatment to improve discourse production.

2) *How does word retrieval ability at discourse level change during normal aging*?

The rate of discourse changes and how age-related word retrieval changes in discourse affect communicative effectiveness and efficiency are not very well understood. Even though our results showed that CLA was not reliable over a period of months, it may be more reliable over a shorter period of time. There are currently few studies on the test-retest reliability of discourse measures in general, and a better understanding of how discourse changes with aging can inform the design of reliability studies, specifically the appropriate lengths of time intervals in which the constructs in question are likely to be stable.

Appendix A Checklist for Best Practices

Category	Item Number	**Reporting Standard**	Included (mark "x")	Page Number(s)
Information about discourse sample	1	Define "discourse"	x	2
	2	Define "utterance" (or other units, e.g., turn unit)		
	3*	Number of words in sample		
Information about how the discourse sample was collected	4	Describe elicitation task	x	25
	5	Exact instructions used to elicit discourse sample	x	25
Information about the persons included in the collection of the discourse sample	6	Demographic information about primary speaker [the person whose discourse is of interest]	x	20–22
	7	Information about the primary speaker's neurological condition	x	20
Methodology and rater agreement	8	Inter-rater reliability for each analyzed variable/measure		
	9	Reliability statistics used		
	10	Details on the number (percentage) of files used for determining reliability/agreement		
	11*	Reliability (point-to-point agreement) for transcription		
Analysis	12	Type of transcription (e.g., orthographic, phonetics)	x	26
	13	Detailed description of any perceptual rating scale used, including providing a copy of the scale if not previously published		
	14	Details of the annotation system, formal (e.g., CHAT) or informal (created by the clinician/examiner)	x	26–27
	15	Whether transcription was verbatim (e.g., including all behaviors such as fillers) or	x	26

		whether information was excluded in the transcription process.		
	16	Completeness of transcription (full, partial, transcribing errors only)	x	26
	17*	Details of any software used for transcribing/annotating/generating data (e.g., SALT,CLAN, ELAN)	x	27
	18*	Who/what transcribed the sample (by a human, by a machine/software, hybrid human and software)	x	26
Information about the individual discourse variables/behaviors reported	19	What is being used as primary outcome measure(s) (e.g., linguistic information, speech information, etc.)	x	8
	20*	Theoretical rationale for selecting variable/behavior/outcome measure(s)	x	8–12
	21	Operational definition for each variable/behavior/outcome(s)	x	14, 15, 16, 27–28

Appendix B List of Stimuli

Appendix B.1 Complex Scenes

1. Opening scene, pages 1–2 in *Where's Waldo?* (Handford, 1987); hereafter referred to as "Waldo-1".
2. "Campsite", pages 7–8 in *Where's Waldo?* (Handford, 1987).
3. "Airport", pages 11–12 in *Where's Waldo?* (Handford, 1987).
4. "Fun and Games in Ancient Rome," pages 5–6 in *Where's Waldo Now?* (Handford, 1988); hereafter referred to as "Ancient Rome".
5. "Being a Pirate", pages 17–18 in *Where's Waldo Now?* (Handford, 1988).
6. "The Deep Sea Divers", pages 15–16 in *Where's Waldo? The Fantastic Journey* (Handford, 1989); hereafter referred to as "Deep Sea Divers".
7. "The Unfriendly Giants", pages 19–20 in *Where's Waldo? The Fantastic Journey* (Handford, 1989); hereafter referred to as "Unfriendly Giants".
8. "The Wild, Wild West", pages 11–12 in *Where's Waldo in Hollywood?* (Handford, 1993); hereafter referred to as "Wild West".
9. "Dinosaurs, Spacemen, and Ghouls", pages 15–16 in *Where's Waldo in Hollywood?* (Handford, 1993).
10. "Exhibit 12—the Great Portrait Exhibition", pages 17–18 in *Where's Waldo? The Great Picture Hunt!* (Handford, 2006); hereafter referred to as "Exhibit 12".
11. "Picture This", pages 5–6 in *Where's Waldo? The Incredible Paper Chase* (Handford, 2009).

12. “Escape from the Zoo”, pages 1–2 in *Where’s the Penguin?* (Schrey & Whelon, 2017).
13. “Birthday Party Fun”, pages 5–6 in *Where’s the Penguin?* (Schrey & Whelon, 2017).
14. “By the Lake”, pages 9–10 in *Where’s the Penguin?* (Schrey & Whelon, 2017).
15. “Making a Splash”, pages 19–20 in Where’s the Penguin? (Schrey & Whelon, 2017); hereafter referred to as “Make a Splash”.

Appendix B.2 Wordless Picture Book Sections

1. “Arrival 1”: pages 2–6, 9–10, 14 in *the Arrival* (Tan, 2007).
2. “Arrival 2”: pages 21–26 in *the Arrival* (Tan, 2007).
3. “Arrival 3”: pages 37–40 in *the Arrival* (Tan, 2007).
4. “Arrival 4”: pages 51–66 in *the Arrival* (Tan, 2007).
5. “Arrival 5”: pages 75–78, 80–82 in *the Arrival* (Tan, 2007).
6. “Arrival 6”: pages 105–112, 118–120 in *the Arrival* (Tan, 2007).
7. “Journey”: entire book of *Journey* (Becker, 2013).
8. “Quest”: entire book of *Quest* (Becker, 2014).
9. “Return”: entire book of *Return* (Becker, 2016).
10. “Flowerman”: entire book of *The Flower Man* (Ludy, 2015).
11. “Wanderer”: entire book of *The Wanderer* (Van den Ende, 2020).

Appendix B.3 Missing Data

Stimuli	Participants Who Did Not Complete in Phase I	Participants Who Did Not Complete in Phase II
Airport	Participant 24	
Ancient Rome	Participant 24, Participant 33	
Arrival 1	Participant 24, Participant 33	
Arrival 2	Participant 24, Participant 33	
Arrival 3		
Arrival 4	Participant 8 (missing the last two slides), Participant 24	
Arrival 5	Participant 33	
Arrival 6	Participant 24, Participant 33	
Being a Pirate	Participant 24, Participant 33	
Birthday Party Fun	Participant 24	
By the Lake		
Campsite	Participant 24, Participant 33	
Deep Sea Divers	Participant 24	
Dinosaurs, Spacemen, and Ghouls	Participant 24, Participant 33	
Escape from the Zoo	Participant 24	
Exhibit 12	Participant 24	
Flowerman	Participant 24	
Journey	Participant 33	
Make a Splash	Participant 24	
Picture This	Participant 24	
Quest	Participant 24, Participant 33	
Return	Participant 24, Participant 33	
Unfriendly Giants	Participant 24, Participant 33	Participant 10
Waldo 1		
Wanderer	Participant 24	
Wild West	Participant 24	

Appendix C Spearman's ρ Values Between Percent Agreements and Speaking Durations for Each Participant

Participant	Spearman's ρ	*p*-Value	Significant Correlation?
Participant30	0.715971	3.91E-05	TRUE
Participant15	0.573588	0.002188	TRUE
Participant29	0.568325	0.002453	TRUE
Participant12	0.557436	0.003091	TRUE
Participant1	0.555847	0.003195	TRUE
Participant23	0.522778	0.006143	TRUE
Participant5	0.511899	0.007512	TRUE
Participant14	0.506892	0.008223	TRUE
Participant16	0.447927	0.021747	TRUE
Participant10	0.440517	0.024302	TRUE
Participant25	0.434586	0.026517	TRUE
Participant21	0.417359	0.033891	TRUE
Participant18	0.386744	0.050969	FALSE
Participant9	0.191885	0.347694	FALSE
Participant19	0.143484	0.48438	FALSE

Appendix D Detailed Lexical Properties of Core Noun Checklists

Lemmas are sorted from having the highest to lowest counts for each stimulus.

Airport

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
plane	4	18	0.002664	0.049984	4.95	3.23	646.724**	-1.02999	3395.707	–
luggage	5	2*	0.0023166*	0.077347*	7.63	2.57	630.875**	-0.25741	935.807	–
airport	6	1	0.003946	0.056278	6.84	2.99	555.857**	-0.51531	2183.614	–
airplane	6	1	0.002875	0.047821	3.94	2.55	800	-0.41645	1496.553	13
truck	4	12	0.005329	0.058921	3.79	3.19	990	-0.67961	3135.564	13
flag	4	13	0.001145	0.03702	5.33	2.7	847	-0.46895	3733.463	–
bus	3	39	0.005681	0.071321	3.85	3.19	771	-1.13117	3370.524	17
baggage	5	3	0.001212	0.078925	7.52	2.35	674.462**	-0.1233	1172.264	–
fire	3	24	0.001112	0.066262	3.25	3.6	854	-1.25847	5145.219	–

Ancient Rome

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
lion	4	14	0.009722	0.051314	4.42	2.52	812	-0.58694	2040.712	20
horse	4	18	0.003746	0.077883	4.15	3.2	809	-0.89776	3789.263	19
chariot	6	3	0.005142	0.091399	9.11	1.87	742.667**	0.539962	246.9172	–
arena	5	8	0.002817	0.048152	9.57	2.07	645**	0.25002	2079.953	–
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.15763	5746.324	–
gladiator	8	1	0.002513	0.025858	8.56	1.48	627.111**	0.710699	436.8925	–
wall	3	38	0.002443	0.073465	3.79	3.32	545.143**	-1.52572	4989.163	14
ring	3	42	0.012477	0.05619	4.53	3.32	785	-1.05925	4288.007	–
stand	5	13	0.008171	0.065166	4.39	3.69	627.185**	-1.1868	5250.861	–

Arrival 1

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–
suitcase	6	0*	0.0013413*	0.059088*	8.22	2.57	924	0.401498	765.4584	–
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.15763	5746.324	–
girl	3	23	0.000886	0.052661	4	3.79	877	-1.18618	5028.169	–
train	4	21	0.003847	0.056708	4	3.25	838	-0.89683	4630.31	15
hat	3	49*	0.0023175*	0.090966*	3.33	3.24	692	-1.3458	4138.674	–
child	4	10	0.001303	0.066115	5.15	3.49	588.852**	-1.18741	4563.092	–
woman	5	5*	0.007815*	0.077949*	4.95	3.77	1130	-0.38891	4387.85	–
bird	3	35*	0.0018231*	0.065215*	3.52	3.08	1527	-0.01727	3411.473	–
family	6	3	0.002559	0.057543	3.38	3.73	608.393**	-1.24579	5167.532	–

paper	4	19*	0.0012915*	0.05047*	4	3.42	931	-0.82548	4707.389	–
photo	4	5	0.000777	0.042059	5.16	2.89	593.074**	-0.94505	4076.587	–
father	4	8	0.000986	0.057659	4.11	3.74	659.037**	-1.33715	3943.222	–
spoon	4	16	0.001516	0.049914	2.5	2.43	777	-0.89162	2144.56	19
mother	4	12	0.002606	0.054375	2.63	3.74	565.607**	-1.69761	4387.983	–
origami	7	0*	0.0028236*	0.040463*	9.24	1.34	–	0.1741884†	259.0542	–
bowl	3	39	0.001297	0.058653	4.26	2.83	828	-0.84443	2951.651	14
clock	4	19	0.001549	0.050019	4.42	3.21	776	-0.91008	3368.702	14
table	4	11	0.004924	0.072937	4.39	3.46	848	-0.90562	4753.778	11
ship	3	29*	0.0010862*	0.051638*	5.33	3.07	645.222**	-1.04865	4171.154	13
row	2	50	0.002269	0.044894	7	2.95	545.111**	-1.0356	4074.853	–
wall	3	38	0.002443	0.073465	3.79	3.32	545.143**	-1.52572	4989.163	14
pot	3	43	0.002003	0.074287	5.95	2.88	1532	0.432844	3513.184	10
cup	3	26	0.005181	0.059648	3.57	3.2	923	-0.95181	4149.761	21

Arrival 2

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
boat	3	37*	0.0007957*	0.077222*	3.84	3.16	982	-0.80824	3963.593	10
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.15763	5746.324	–
ship	3	29*	0.0010862*	0.051638*	5.33	3.07	645.222**	-1.04865	4171.154	13
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
bird	3	35*	0.0018231*	0.065215*	3.52	3.08	1527	-0.01727	3411.473	–
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–
suitcase	6	0*	0.0013413*	0.059088*	8.22	2.57	924	0.401498	765.4584	–
hat	3	49*	0.0023175*	0.090966*	3.33	3.24	692	-1.3458	4138.674	–
city	4	17*	0.0046549*	0.087785*	6.56	3.52	1165	-0.13379	4825.942	–
woman	5	5*	0.007815*	0.077949*	4.95	3.77	1130	-0.38891	4387.85	–
statue	5	3*	0.0043466*	0.038801*	7.55	2.5	1234	0.651788	1893.516	–

book	3	28*	0.0008152*	0.055528*	3.68	3.51	656	-1.45222	5009.297	16
luggage	5	2*	0.0023166*	0.077347*	7.63	2.57	630.875**	-0.25741	935.807	–
sky	3	9*	0.0015711*	0.026981*	4.17	3.09	634.679**	-1.2461	4002.131	–
paper	4	19*	0.0012915*	0.05047*	4	3.42	931	-0.82548	4707.389	–
head	3	15*	0.0013972*	0.065609*	3.42	3.8	603.036**	-1.6831	5478.092	–
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
origami	7	0*	0.0028236*	0.040463*	9.24	1.34	–	0.1741884†	259.0542	–
person	4	4*	0.0092245*	0.089649*	4.67	3.67	547.107**	-1.39166	5611.322	–
creature	5	4*	0.003673*	0.055428*	7.32	2.82	646.391**	-0.37867	2124.134	–
line	3	44*	0.0029258*	0.080569*	4.85	3.66	614.607**	-1.39671	5384.215	–
arm	3	15*	0.0050179*	0.030361*	3.26	3.29	910	-1.05277	4290.681	–
deck	3	36*	0.0023342*	0.074852*	6.45	2.8	589.071**	-0.8583	2491.307	–
shoulder	5	8*	0.0019002*	0.044223*	4.5	2.99	1118	-0.16921	3428.435	–

Arrival 3

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.158	5746.324	–
hat	3	49*	0.0023175*	0.090966*	3.33	3.24	692	-1.3458	4138.674	–
woman	5	5*	0.007815*	0.077949*	4.95	3.77	1130	-0.38891	4387.85	–
book	3	28*	0.0008152*	0.055528*	3.68	3.51	656	-1.45222	5009.297	16
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
street	5	9	0.008617	0.078273	4.58	3.53	655.318**	-1.05405	4403.297	–
suitcase	6	0*	0.0013413*	0.059088*	8.22	2.57	924	0.401498	765.4584	–
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–
egg	2	27	0.00045	0.026881	3.89	2.86	883	-0.96615	3251.606	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
ground	5	15	0.003525	0.052857	4.89	3.36	578.429**	-1.05503	4839.127	–
animal	5	3	0.00742	0.069124	2.89	3.14	556.154**	-1.30202	3978.768	–
boy	2	23	0.000373	0.028192	3.67	3.79	964	-1.24274	4942.75	–

cart	4	24	0.004373	0.094986	6.16	2.51	608.889**	-0.62304	2139.103	11
newspaper	7	1	0.001108	0.028422	5.78	2.88	638.296**	-0.37313	1793.269	–
basket	6	6	0.004274	0.100872	5.67	2.63	840	-0.12895	2101.748	9
light	3	39	0.002793	0.087139	4.05	3.59	578.321**	-1.54312	5349.571	–
guy	2	34	0.000392	0.053112	4.11	3.82	571.393**	-1.77032	5556.939	–
creature	5	4*	0.003673*	0.055428*	7.32	2.82	646.391**	-0.37867	2124.134	–
cat	3	45	0.001803	0.085269	3.68	3.14	767	-1.14305	4513.737	20
clock	4	19	0.001549	0.050019	4.42	3.21	776	-0.91008	3368.702	14
map	3	32	0.001682	0.049749	5.6	2.91	847	-0.64656	3868.727	–
bed	3	40	0.002067	0.084996	2.89	3.61	706	-1.53572	4755.127	17
head	3	15*	0.0013972*	0.065609*	3.42	3.8	603.036**	-1.6831	5478.092	–
barber	5	14	0.002276	0.05703	6.68	2.18	586.667**	-0.31294	649.5971	–

Arrival 4

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.158	5746.324	–
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
boy	2	23	0.000373	0.028192	3.67	3.79	964	-1.24274	4942.75	–
book	3	28*	0.0008152*	0.055528*	3.68	3.51	656	-1.45222	5009.297	16
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–
creature	5	4*	0.003673*	0.055428*	7.32	2.82	646.391**	-0.37867	2124.134	–
animal	5	3	0.00742	0.069124	2.89	3.14	556.154**	-1.30202	3978.768	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
woman	5	5*	0.007815*	0.077949*	4.95	3.77	1130	-0.38891	4387.85	–
map	3	32	0.001682	0.049749	5.6	2.91	847	-0.64656	3868.727	–
boat	3	37*	0.0007957*	0.077222*	3.84	3.16	982	-0.80824	3963.593	10
hat	3	49*	0.0023175*	0.090966*	3.33	3.24	692	-1.3458	4138.674	–
basket	6	6	0.004274	0.100872	5.67	2.63	840	-0.12895	2101.748	9
ticket	5	6	0.005845	0.124183	5.32	3.12	675**	-0.75357	3675.284	–

drawer	4	9	0.003791	0.039268	5.37	2.67	994	-0.23336	1525.459	–
city	4	17*	0.0046549*	0.087785*	6.56	3.52	1165	-0.13379	4825.942	–
tail	3	37	0.001528	0.06275	3.7	2.91	1432	-0.06459	3137.327	–
father	4	8	0.000986	0.057659	4.11	3.74	659.037**	-1.33715	3943.222	–
girl	3	23	0.000886	0.052661	4	3.79	877	-1.18618	5028.169	–
train	4	21	0.003847	0.056708	4	3.25	838	-0.89683	4630.31	15
fish	3	18	0.001785	0.057044	4.05	3.23	777	-1.10792	3983.347	–
door	3	40	0.004509	0.062343	3.05	3.72	719	-1.53546	4771.281	13
shovel	4	5	0.005247	0.074013	6.28	2.4	858	-0.19284	1308.892	15
building	6	8	0.006848	0.108056	6.16	3.38	620.808**	-0.67354	4830.169	15
ship	3	29*	0.0010862*	0.051638*	5.33	3.07	645.222**	-1.04865	4171.154	13

Arrival 5

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.158	5746.324	–
woman	5	5*	0.007815*	0.077949*	4.95	3.77	1130	-0.38891	4387.85	–
boy	2	23	0.000373	0.028192	3.67	3.79	964	-1.24274	4942.75	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
hat	3	49*	0.0023175*	0.090966*	3.33	3.24	692	-1.3458	4138.674	–
creature	5	4*	0.003673*	0.055428*	7.32	2.82	646.391**	-0.37867	2124.134	–
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
pet	3	37	0.0024	0.087614	4.05	2.8	606.964**	-1.19178	3539.392	–
food	3	23	0.000524	0.061991	3.25	3.56	571.185**	-1.66196	4946.764	–
animal	5	3	0.00742	0.069124	2.89	3.14	556.154**	-1.30202	3978.768	–
cat	3	45	0.001803	0.085269	3.68	3.14	767	-1.14305	4513.737	20
table	4	11	0.004924	0.072937	4.39	3.46	848	-0.90562	4753.778	11
basket	6	6	0.004274	0.100872	5.67	2.63	840	-0.12895	2101.748	9
boat	3	37*	0.0007957*	0.077222*	3.84	3.16	982	-0.80824	3963.593	10

bowl	3	39	0.001297	0.058653	4.26	2.83	828	-0.84443	2951.651	14
building	6	8	0.006848	0.108056	6.16	3.38	620.808**	-0.67354	4830.169	15
pot	3	43	0.002003	0.074287	5.95	2.88	1532	0.432844	3513.184	10
bird	3	35*	0.0018231*	0.065215*	3.52	3.08	1527	-0.01727	3411.473	–
suit	3	33	0.001074	0.075109	6.67	3.29	673.444**	-0.89185	4311.902	–
city	4	17*	0.0046549*	0.087785*	6.56	3.52	1165	-0.13379	4825.942	–
guy	2	34	0.000392	0.053112	4.11	3.82	571.393**	-1.77032	5556.939	–
instrument	10	1	0.011687	0.138116	6.94	2.5	634.808**	0.352285	1695.267	–
glass	4	13	0.001797	0.056756	4.47	3.27	942	-0.68003	4258.68	–
father	4	8	0.000986	0.057659	4.11	3.74	659.037**	-1.33715	3943.222	–

Arrival 6

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.15763	5746.324	–
letter	4	26	0.00346	0.072988	4.74	3.25	988	-0.56353	3966.63	–
flower	4	7	0.000713	0.055567	3.11	2.8	754	-0.97887	2168.826	–
bird	3	35*	0.0018231*	0.065215*	3.52	3.08	1527	-0.01727	3411.473	–
creature	5	4*	0.003673*	0.055428*	7.32	2.82	646.391**	-0.37867	2124.134	–
pot	3	43	0.002003	0.074287	5.95	2.88	1532	0.432844	3513.184	10
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–
envelope	7	1	0.004263	0.050243	6.25	2.51	803	0.086136	1523.271	11
animal	5	3	0.00742	0.069124	2.89	3.14	556.154**	-1.30202	3978.768	–
ground	5	15	0.003525	0.052857	4.89	3.36	578.429**	-1.05503	4839.127	–
window	5	8	0.009407	0.06417	4.74	3.39	822	-0.72829	4594.18	–
paper	4	19*	0.0012915*	0.05047*	4	3.42	931	-0.82548	4707.389	–
girl	3	23	0.000886	0.052661	4	3.79	877	-1.18618	5028.169	–
hat	3	49*	0.0023175*	0.090966*	3.33	3.24	692	-1.3458	4138.674	–
family	6	3	0.002559	0.057543	3.38	3.73	608.393**	-1.24579	5167.532	–

leaf	3	27	0.004008	0.048813	4.6	2.34	848	-0.5722	1972.37	–
building	6	8	0.006848	0.108056	6.16	3.38	620.808**	-0.67354	4830.169	15
origami	7	0*	0.0028236*	0.040463*	9.24	1.34	–	0.1741884†	259.0542	–
bed	3	40	0.002067	0.084996	2.89	3.61	706	-1.53572	4755.127	17
pet	3	37	0.0024	0.087614	4.05	2.8	606.964**	-1.19178	3539.392	–
mailbox	7	0	0.001968	0.041766	6.11	2.2	856	0.264892	902.4789	–
piece	3	36	0.001674	0.05409	4.89	3.55	615.296**	-1.34668	5149.263	–
shape	3	20	0.000357	0.038342	4.47	3.08	616.929**	-1.22349	4473.82	–
mouse	3	22	0.000241	0.045364	4.94	2.6	954	-0.46577	2845.936	19

Being a Pirate

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
boat	3	37*	0.0007957*	0.077222*	3.84	3.16	982	-0.8082	3963.593	10
ship	3	29*	0.0010862*	0.051638*	5.33	3.07	645.222**	-1.04865	4171.154	13
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
pirate	5	4	0.003835	0.093564	6.5	2.21	1186	0.536542	1977.962	–
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
cannon	5	7	0.009919	0.100493	7.9	2.41	1159	0.628328	1850.127	17
flag	4	13	0.001145	0.03702	5.33	2.7	847	-0.46895	3733.463	–
shore	3	29*	0.003208*	0.052366*	6.93	2.66	642**	-0.653116	1772.181	–
palm tree	palm - 4	3*	0.0012794*	0.0544935*	7.44	2.6395	914	–	9.20439	–
	tree - 3	23*	0.0081493*	0.0330463*		3.2103				
mast	4	28	0.006553	0.080747	10.59	1.76	589.333**	0.305732	389.9474	–

Birthday Party Fun

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
costume	6	3	0.004364	0.055492	4.17	2.6	638.519**	-0.64081	2243.836	–
party	5	10	0.003914	0.064939	4.58	3.57	558.036**	-1.21386	5205.991	–
birthday	5	1	0.000762	0.033705	2.85	3.22	597**	-1.27877	3906.063	–
table	4	11	0.004924	0.072937	4.39	3.46	848	-0.90562	4753.778	11
kid	3	32	0.004403	0.08664	4.28	3.67	566.333**	-1.55766	4813.994	–
pool	3	29	0.000457	0.045932	5.15	3.08	879	-0.7331	4191.752	–
cake	3	36	0.001788	0.06481	3.26	3.08	789	-1.14998	4006.901	17
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
food	3	23	0.000524	0.061991	3.25	3.56	571.185**	-1.66196	4946.764	–

By the Lake

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
boat	3	37*	0.0007957*	0.077222*	3.84	3.16	982	-0.80824	3963.593	10
lake	3	38	0.002408	0.065655	4.61	2.9	530.815**	-1.25973	2716.593	–
beach	3	29	0.00123	0.05021	4.8	3.11	623.64**	-1.17578	3598.062	–
kid	3	32	0.004403	0.08664	4.28	3.67	566.333**	-1.55766	4813.994	–
penguin	7	1	0.005278	0.076129	5.68	1.79	897	0.421464	938.5638	18
dock	3	38	0.001197	0.061524	8.22	2.49	569.593**	-0.50045	1231.273	–
rock	3	43	0.001574	0.058183	3.22	3.28	912	-1.0519	4793.697	12
kite	3	25	0.001543	0.086294	4.58	1.89	796	-0.47632	654.1918	14

Campsite

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
tent	4	27	0.006695	0.104799	5.16	2.65	744	-0.62754	1733.46	11
river	4	17	0.004155	0.055513	4.9	3.08	600.071**	-1.04797	2997.933	–
camper	5	13	0.001999	0.056081	6.67	1.69	682.857**	0.019704	590.4926	–
bull	3	23	0.000469	0.056562	4.53	2.82	569.75**	-1.18275	2483.82	12
area	4	9	0.006179	0.038945	6.58	3.33	570.852**	-0.93703	5125.241	–
RV	–	–	–	–	–	–	–	–	–	–
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.15763	5746.324	–
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
bridge	4	12	0.003538	0.058939	5.58	3.02	860	-0.5373	3552.148	14

Deep Sea Divers

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
fish	3	18	0.001785	0.057044	4.05	3.23	777	-1.108	3983.347	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
boat	3	37*	0.0007957*	0.077222*	3.84	3.16	982	-0.80824	3963.593	10
sea	2	56	0.002298	0.095836	4.74	3.13	713.917**	-1.19481	3790.854	–
mermaid	5	2	0.001162	0.051715	5.68	1.87	731.296**	-0.12739	635.4726	–
rock	3	43	0.001574	0.058183	3.22	3.28	912	-1.0519	4793.697	12
bottom	5	4	0.00477	0.077854	5.28	3.25	601.786**	-0.9189	5095.262	–
net	3	32	0.001978	0.09056	7	2.73	1007	-0.12906	4103.667	–
diver	4	12	0.001557	0.048729	6.89	1.73	710.556**	-0.05789	393.2787	–

Dinosaurs, Spacemen, and Ghouls

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
dinosaur	7	1	0.004223	0.05244	4.91	2.07	1013	0.3684	1287.377	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
caveman	6	2*	0.0066149*	0.0715621*	6.28	1.69	654.308**	0.0548945	466.2293	–
movie	4	7	0.000627	0.051039	3.56	3.32	605.885**	-1.3342	4615.803	–
light	3	39	0.002793	0.087139	4.05	3.59	578.321**	-1.54312	5349.571	–
space	4	11	0.002467	0.067227	5.67	3.22	688.926**	-0.85553	5164.963	–
volcano	7	1	0.00121	0.036268	6.74	2.04	1063	0.728726	934.8191	–
spaceman	7	0*	0.0064737*	0.0785675*	7.63	1.4	–	-0.1856937†	90.03531	–
set	3	40	0.003381	0.098222	6	3.71	694.679**	-1.1251	5626.325	–
area	4	9	0.006179	0.038945	6.58	3.33	570.852**	-0.93703	5125.241	–

Escape from the Zoo

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
zoo	2	24	4.51E-05	0.047631	3.75	2.59	662.571**	-1.20826	1460.659	–
animal	5	3	0.00742	0.069124	2.89	3.14	556.154**	-1.30202	3978.768	–
penguin	7	1	0.005278	0.076129	5.68	1.79	897	0.421464	938.5638	18
rhinoceros	9	0	0.002589	0.070133	6	1.48	1055	1.09683	120.3628	–
enclosure	7	2	0.004872	0.039067	9.72	1.26	716.923**	0.96782	441.2403	–
lion	4	14	0.009722	0.051314	4.42	2.52	812	-0.58694	2040.712	20
ice cream	ice - 3	24*	0.0009192*	0.0196899*	3.33	0.0196899	–	3.2956	289.254	–
	cream - 4	21*	0.0049666*	0.0551052*		0.0551052		3.129		
bird	3	35*	0.0018231*	0.065215*	3.52	3.08	1527	-0.01727	3411.473	–
giraffe	5	1	0.000506	0.018996	5	1.76	783	-0.10975	547.3671	14

Exhibit 12

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
portrait	7	1	0.005539	0.09993	7.61	2.29	679.346**	0.192901	1320.735	–
wall	3	38	0.002443	0.073465	3.79	3.32	545.143**	-1.52572	4989.163	14
level	4	6	0.004624	0.081562	6	3.2	567.296**	-0.97847	5324.137	9
floor	4	12	0.003654	0.039378	4.44	3.44	1463	0.021126	4546.877	–
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.15763	5746.324	–
ladder	4	15	0.002012	0.058392	4.4	2.5	988	-0.32128	2066.692	–
painting	6	8	0.007614	0.105678	5.26	2.81	646.68**	-0.54733	2859.969	–
frame	4	13	0.002441	0.050782	7.67	2.72	630.885**	-0.44605	4087.228	–

Flowerman

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
bird	3	35*	0.0018231*	0.065215*	3.52	3.08	1527	-0.01727	3411.473	–
girl	3	23	0.000886	0.052661	4	3.79	877	-1.18618	5028.169	–
boat	3	37*	0.0007957*	0.077222*	3.84	3.16	982	-0.80824	3963.593	10
door	3	40	0.004509	0.062343	3.05	3.72	719	-1.53546	4771.281	13
boy	2	23	0.000373	0.028192	3.67	3.79	964	-1.24274	4942.75	–
cage	3	19	0.000919	0.042846	5.06	2.75	987	-0.45752	2342.943	13
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
crayon	5	2	0.002356	0.056483	3.2	1.32	599.962**	-0.47898	475.0741	13
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–

marker	5	11	0.002436	0.059452	4.89	2.21	562.333**	-0.62923	1608.387	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
city	4	17*	0.0046549*	0.087785*	6.56	3.52	1165	-0.13379	4825.942	–
guard	4	19	0.003887	0.076359	6.25	3.2	596.107**	-0.89827	3728.524	–
castle	4	13	0.006054	0.089189	5.8	2.58	893	-0.28335	2236.356	–
circle	4	10	0.005798	0.098794	3.67	2.84	635.143**	-1.08666	4145.218	–
light	3	39	0.002793	0.087139	4.05	3.59	578.321**	-1.54312	5349.571	–
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
room	3	32	0.001491	0.04533	4.22	3.8	568.423**	-1.6144	5309.511	–
tree	3	23	0.008149	0.033046	3.57	3.21	796	-1.14394	4223.873	–
wall	3	38	0.002443	0.073465	3.79	3.32	545.143**	-1.52572	4989.163	14
air	2	43	0.005887	0.049423	3.94	3.52	587.538**	-1.65459	4997.552	–
sky	3	9*	0.0015711*	0.026981*	4.17	3.09	634.679**	-1.2461	4002.131	–
carpet	6	2	0.003814	0.098987	6.05	2.59	604.44**	-0.40541	2038.034	17
balloon	5	7	0.004718	0.06008	4.37	2.42	702	-0.58236	1508.147	18
magic	5	2	0.00236	0.076276	5.52	3.09	645.731**	-0.75522	4435.947	–

Journey

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
bird	3	35*	0.0018231*	0.065215*	3.52	3.08	1527	-0.01727	3411.473	–
girl	3	23	0.000886	0.052661	4	3.79	877	-1.18618	5028.169	–
boat	3	37*	0.0007957*	0.077222*	3.84	3.16	982	-0.80824	3963.593	10
door	3	40	0.004509	0.062343	3.05	3.72	719	-1.53546	4771.281	13
boy	2	23	0.000373	0.028192	3.67	3.79	964	-1.24274	4942.75	–
cage	3	19	0.000919	0.042846	5.06	2.75	987	-0.45752	2342.943	13
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
crayon	5	2	0.002356	0.056483	3.2	1.32	599.962**	-0.47898	475.0741	13
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–
marker	5	11	0.002436	0.059452	4.89	2.21	562.333**	-0.62923	1608.387	–

people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
city	4	17*	0.0046549*	0.087785*	6.56	3.52	1165	-0.13379	4825.942	–
guard	4	19	0.003887	0.076359	6.25	3.2	596.107**	-0.89827	3728.524	–
castle	4	13	0.006054	0.089189	5.8	2.58	893	-0.28335	2236.356	–
circle	4	10	0.005798	0.098794	3.67	2.84	635.143**	-1.08666	4145.218	–
light	3	39	0.002793	0.087139	4.05	3.59	578.321**	-1.54312	5349.571	–
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
room	3	32	0.001491	0.04533	4.22	3.8	568.423**	-1.6144	5309.511	–
tree	3	23	0.008149	0.033046	3.57	3.21	796	-1.14394	4223.873	–
wall	3	38	0.002443	0.073465	3.79	3.32	545.143**	-1.52572	4989.163	14
air	2	43	0.005887	0.049423	3.94	3.52	587.538**	-1.65459	4997.552	–
sky	3	9*	0.0015711*	0.026981*	4.17	3.09	634.679**	-1.2461	4002.131	–
carpet	6	2	0.003814	0.098987	6.05	2.59	604.44**	-0.40541	2038.034	17
balloon	5	7	0.004718	0.06008	4.37	2.42	702	-0.58236	1508.147	18
magic	5	2	0.00236	0.076276	5.52	3.09	645.731**	-0.75522	4435.947	–

Make a Splash

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
water	4	18*	0.0031341*	0.064564*	2.37	3.62	569.704**	-1.68362	5152.861	–
slide	4	20	0.002118	0.082405	4.79	2.76	1003	-0.34221	3526.904	–
kid	3	32	0.004403	0.08664	4.28	3.67	566.333**	-1.55766	4813.994	–
ship	3	29*	0.0010862*	0.051638*	5.33	3.07	645.222**	-1.04865	4171.154	13
pool	3	29	0.000457	0.045932	5.15	3.08	879	-0.7331	4191.752	–
tube	3	2	0.001505	0.019881	5.5	2.66	605.808**	-0.92125	2258.428	–
park	4	22	0.003169	0.060532	4.47	3.27	558.571**	-1.24827	4215.573	–
river	4	17	0.004155	0.055513	4.9	3.08	600.071**	-1.04797	2997.933	–
lifeguard	7	1	0.002493	0.032299	5.63	1.75	602.556**	-0.00676	253.3316	–

Picture This

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
man	3	48*	0.0044106*	0.077868*	3.11	3.91	1018	-1.15763	5746.324	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
person	4	4*	0.0092245*	0.089649*	4.67	3.67	547.107**	-1.39166	5611.322	–
head	3	15*	0.0013972*	0.065609*	3.42	3.8	603.036**	-1.6831	5478.092	–
portrait	7	1	0.005539	0.09993	7.61	2.29	679.346**	0.192901	1320.735	–
hand	4	22*	0.006294*	0.09054*	2.74	3.75	721	-1.45473	5472.223	–
guy	2	34	0.000392	0.053112	4.11	3.82	571.393**	-1.77032	5556.939	–
hat	3	49*	0.0023175*	0.090966*	3.33	3.24	692	-1.3458	4138.674	–
caveman	6	2*	0.0066149*	0.0715621*	6.28	1.6902	654.308**	0.0548945	466.2293	–

Quest

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
bird	3	35	0.001823	0.065215	3.52	3.08	1527	-0.01727	3411.473	–
boy	2	23	0.000373	0.028192	3.67	3.79	964	-1.24274	4942.75	–
girl	3	23	0.000886	0.052661	4	3.79	877	-1.18618	5028.169	–
king	3	36	0.010586	0.061718	5.42	3.23	898	-0.72309	4494.921	–
child	4	10	0.001303	0.066115	5.15	3.49	588.852**	-1.18741	4563.092	–
door	3	40	0.004509	0.062343	3.05	3.72	719	-1.53546	4771.281	13
bridge	4	12	0.003538	0.058939	5.58	3.02	860	-0.5373	3552.148	14
water	4	12	0.002964	0.062732	2.37	3.62	569.704**	-1.68362	5152.861	–
rainbow	5	1	0.001648	0.03992	4.26	2.37	1014	-0.11692	2280.845	–
boat	3	37	0.000796	0.077222	3.84	3.16	982	-0.80824	3963.593	10
crayon	5	2	0.002356	0.056483	3.2	1.32	599.962**	-0.47898	475.0741	13

guard	4	19	0.003887	0.076359	6.25	3.2	596.107**	-0.89827	3728.524	–
circle	4	10	0.005798	0.098794	3.67	2.84	635.143**	-1.08666	4145.218	–
city	4	17	0.004655	0.087785	6.56	3.52	1165	-0.13379	4825.942	–
soldier	5	2	0.000898	0.060451	6.94	2.92	1324	0.529448	2203.005	–
man	3	48	0.004411	0.077868	3.11	3.91	1018	-1.15763	5746.324	–
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
kid	3	32	0.004403	0.08664	4.28	3.67	566.333**	-1.55766	4813.994	–
hand	4	22	0.006294	0.09054	2.74	3.75	721	-1.45473	5472.223	–
castle	4	13	0.006054	0.089189	5.8	2.58	893	-0.28335	2236.356	–
map	3	32	0.001682	0.049749	5.6	2.91	847	-0.64656	3868.727	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
rhinoceros	9	0	0.002589	0.070133	6	1.48	1055	1.09683	120.3628	–
building	6	0*	0.0045517*	0.051766*	6.16	3.38	620.808**	-0.67354	4830.169	15
statue	5	3	0.004347	0.038801	7.55	2.5	1234	0.651788	1893.516	–

Return

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
man	3	48	0.004411	0.077868	3.11	3.91	1018	-1.158	5746.324	–
bird	3	35	0.001823	0.065215	3.52	3.08	1527	-0.01727	3411.473	–
boy	2	23	0.000373	0.028192	3.67	3.79	964	-1.24274	4942.75	–
boat	3	37	0.000796	0.077222	3.84	3.16	982	-0.80824	3963.593	10
king	3	36	0.010586	0.061718	5.42	3.23	898	-0.72309	4494.921	–
girl	3	23	0.000886	0.052661	4	3.79	877	-1.18618	5028.169	–
door	3	40	0.004509	0.062343	3.05	3.72	719	-1.53546	4771.281	13
father	4	8	0.000986	0.057659	4.11	3.74	659.037**	-1.33715	3943.222	–
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
child	4	10	0.001303	0.066115	5.15	3.49	588.852**	-1.18741	4563.092	–
box	4	31	0.00301	0.075723	4.3	3.37	753	-1.02472	5077.72	14
cave	3	27	0.001211	0.051009	6.74	2.56	589.607**	-0.72016	2611.431	–

kite	3	25	0.001543	0.086294	4.58	1.89	796	-0.47632	654.1918	14
water	4	12	0.002964	0.062732	2.37	3.62	569.704**	-1.68362	5152.861	–
hand	4	22	0.006294	0.09054	2.74	3.75	721	-1.45473	5472.223	–
room	3	32	0.001491	0.04533	4.22	3.8	568.423**	-1.6144	5309.511	–
crayon	5	2	0.002356	0.056483	3.2	1.32	599.962**	-0.47898	475.0741	13
wall	3	38	0.002443	0.073465	3.79	3.32	545.143**	-1.52572	4989.163	14
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
soldier	5	2	0.000898	0.060451	6.94	2.92	1324	0.529448	2203.005	–
dragon	6	1*	0.0063613*	0.078838*	5.58	2.44	891	0.007422	2972.385	–
submarine	8	1	0.002319	0.060329	7.94	2.15	1144	1.121758	624.8565	13
castle	4	13	0.006054	0.089189	5.8	2.58	893	-0.28335	2236.356	–
light	3	39	0.002793	0.087139	4.05	3.59	578.321**	-1.54312	5349.571	–
ball	3	38	0.001676	0.06495	2.9	3.32	886	-1.15407	4636.898	9

Unfriendly Giants

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
giant	5	1	0.008932	0.048096	4.72	2.94	625.815**	-0.8461	4129.61	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
chess	3	20	0.002553	0.058029	5.37	2.28	630.038**	-0.75626	1545.032	–
house	3	15	0.000444	0.051893	3.16	3.78	755	-1.48907	5243.384	16
tree	3	23	0.008149	0.033046	3.57	3.21	796	-1.14394	4223.873	–
windmill	7	2	0.007505	0.078139	5.32	1.6	1252	0.967864	285.7732	–
castle	4	13	0.006054	0.089189	5.8	2.58	893	-0.28335	2236.356	–
woman	5	5	0.007815	0.077949	4.95	3.77	1130	-0.38891	4387.85	–
head	3	39	0.002285	0.091269	3.42	3.8	603.036**	-1.6831	5478.092	–
water	4	12	0.002964	0.062732	2.37	3.62	569.704**	-1.68362	5152.861	–

Waldo 1

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.39546	–
building	6	0*	0.0045517*	0.051766*	6.16	3.38	620.808**	-0.6735425	4830.16913	15
car	3	38	0.00439055	0.06374957	3.37	3.71	751	-1.436128	4735.95579	16
window	5	8	0.00940743	0.06416956	4.74	3.39	822	-0.728286	4594.17995	–
roof	3	24	0.00103986	0.03041141	5	3.06	1092	-0.432116	3150.57011	–
fountain	5	2	0.00914728	0.09411263	7.17	2.38	990	0.2801	1390.68491	–
street	5	9	0.00861667	0.07827268	4.58	3.53	655.318**	-1.0540487	4403.29687	–
man	3	48	0.0044106	0.0778675	3.11	3.91	1018	-1.157634	5746.32397	–
town	3	20	0.00115455	0.05897149	5.11	3.61	587**	-1.379076	4551.66272	–
triangle	7	1	0.00346763	0.06403083	4.9	2.1	635.115**	-0.2048596	1749.77540	–

Wanderer

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
boat	3	37	0.000796	0.077222	3.84	3.16	982	-0.8082	3963.593	10
water	4	12	0.002964	0.062732	2.37	3.62	569.704**	-1.68362	5152.861	–
fish	3	18	0.001785	0.057044	4.05	3.23	777	-1.10792	3983.347	–
paper	4	19	0.001292	0.05047	4	3.42	931	-0.82548	4707.389	–
creature	5	4	0.003673	0.055428	7.32	2.82	646.391**	-0.37867	2124.134	–
sea	2	56	0.002298	0.095836	4.74	3.13	713.917**	-1.19481	3790.854	–
ship	3	29	0.001086	0.051638	5.33	3.07	645.222**	-1.04865	4171.154	13
hat	3	49	0.002317	0.090966	3.33	3.24	692	-1.3458	4138.674	–
bird	3	35	0.001823	0.065215	3.52	3.08	1527	-0.01727	3411.473	–
head	3	39	0.002285	0.091269	3.42	3.8	603.036**	-1.6831	5478.092	–
light	3	39	0.002793	0.087139	4.05	3.59	578.321**	-1.54312	5349.571	–

picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.61482	5912.395	–
eye	1	41	0	0.158505	3.75	3.51	703	-1.64406	5105.27	–
whale	4	40	0.001418	0.073784	5.47	2.37	1075	0.01898	1582.732	12
sky	3	9	0.001571	0.026981	4.17	3.09	634.679**	-1.2461	4002.131	–
origami	7	0*	0.002824*	0.040463*	9.24	1.34	–	0.1741884†	259.0542	–
man	3	48	0.004411	0.077868	3.11	3.91	1018	-1.15763	5746.324	–
submarine	8	1	0.002319	0.060329	7.94	2.15	1144	1.121758	624.8565	13
seahorse	6	0*	0.0024513*	0.0583409*	5.83	0.903	1157	1.040101	93.27669	–
hand	4	22	0.006294	0.09054	2.74	3.75	721	-1.45473	5472.223	–
ocean	3	10	0.010295	0.027598	4.74	2.94	559.321**	-1.21363	3123.164	–
tree	3	23	0.008149	0.033046	3.57	3.21	796	-1.14394	4223.873	–
person	4	4	0.009224	0.089649	4.67	3.67	547.107**	-1.39166	5611.322	–
animal	5	3	0.00742	0.069124	2.89	3.14	556.154**	-1.30202	3978.768	–

Wild West

Lemma	Number of Phonemes	Unstressed Phonological Neighborhood Density	Unstressed Biphoneme Probability	Unstressed Propositional Probability	Age of Acquisition (years)	logCD	Response Time (ms)	Item Difficulty	Contextual Diversity	Num. of Features
people	4	5*	0.0047155*	0.080626*	3.52	3.9	573.6**	-1.615	5912.395	–
saloon	5	6*	0.0054835*	0.070907*	9	2.16	676.308**	0.175728	220.8596	–
horse	4	19*	0.0032454*	0.079715*	4.15	3.2	809	-0.89776	3789.263	19
picture	5	8*	0.0024713*	0.075253*	4.05	3.5	1022	-0.5784	5322.422	–
hotel	5	2*	0.0024567*	0.043681*	6.05	3.31	581.963**	-0.85624	3137.238	–
building	6	0*	0.0045517*	0.051766*	6.16	3.38	620.808**	-0.67354	4830.169	15
jail	3	18*	0.0004288*	0.045882*	5.74	3.23	652.778**	-1.03851	2811.068	–
train	4	21*	0.0038471*	0.056708*	4	3.25	838	-0.89683	4630.31	15
cowboy	4	1*	0.0002781*	0.027676*	4.55	2.68	1341	0.153982	1376.405	–
bank	4	8*	0.0029711*	0.073163*	6.44	3.2	530.893**	-0.96642	4024.032	–

– Values not available in databases or cannot be calculated.

† Item difficulty was calculated using an earlier model developed by Fergadiotis et al. (2015).

* Phonological characteristics were calculated by manually entering the words into the online IPhOD calculator (available at https://www.iphod.com/calculator/V2CalcWords.html) instead of obtained from the downloadable IPhOD data file.

** Response time was obtained from the ELP (Balota et al., 2007; available at https://elexicon.wustl.edu/index.html).

Bibliography

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*, 814–823. http://doi.org/10.1111/j.1467-9280.2006.01787.x.

Alyahya, R. S., Conroy, P., Halai, A. D., & Lambon Ralph, M. A. (2022). An efficient, accurate and clinically-applicable index of content word fluency in aphasia. *Aphasiology*, *36*(8), 921–939. https://doi.org/10.1080/02687038.2021.1923946.

Andreetta, S., Cantagallo, A., & Marini, A. (2012). Narrative discourse in anomic aphasia. *Neuropsychologia*, *50*(8), 1787–1793. https://doi.org/10.1016/j.neuropsychologia.2012.04.003.

Balota, D.A., Yap, M.J., Hutchison, K.A., Cortese, M.J., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B. & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. https://doi.org/10.3758/BF03193014.

Boyle, M., & Coelho, C. A. (1995). Application of semantic feature analysis as a treatment for aphasic dysnomia. *American Journal of Speech-Language Pathology*, *4*(4), 94–98. https://doi.org/10.1044/1058-0360.0404.94.

Boyle, M. (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research*, *57*(3), 966–978. https://doi.org/10.1044/2014_JSLHR-L-13-0171.

Boyle, M. (2015). Stability of word retrieval errors with the AphasiaBank stimuli. *American Journal of Speech-Language Pathology*, *24*(4), S953–S960. https://doi.org/10.1044/2015_AJSLP-14-0152.

Brookshire, R.H. & Nicholas, L.E. (1994a). Speech sample size and test-retest reliability of connected speech measures for adults with aphasia. *Journal of Speech, Language, and Hearing Research*, *37*(2), 399–407. https://doi.org/10.1044/jshr.3702.399.

Brookshire, R. H., & Nicholas, L. E. (1994b). Test-retest stability of measures of connected speech in aphasia. *Clinical Aphasiology*, *22*, 119–133. Retrieved from http://aphasiology.pitt.edu/163/1/22-09.pdf.

Bryant, L., Spencer, E., Ferguson, A., Craig, H., Colyvas, K. &Worrall, L. (2013). Propositional idea density in aphasic discourse. *Aphasiology*, *27*(8), 992–1009. https://doi.org/10.1080/02687038.2013.803514.

Bryant, L., Ferguson, A. & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics*, *30*(7), 489–518. https://doi.org/10.3109/02699206.2016.1145740.

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. https://doi.org/10.1080/02687038.2016.1239013.

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41, 977–990. https://doi.org/10.3758/BRM.41.4.977.

Cahana-Amitay, D., & Jenkins, T. (2018). Working memory and discourse production in people with aphasia. *Journal of Neurolinguistics*, *48*, 90–103. https://doi.org/10.1016/j.jneuroling.2018.04.007.

Cannizzaro, M. S., & Coelho, C. A. (2013). Analysis of narrative discourse structure as an ecologically relevant measure of executive function in adults. *Journal of psycholinguistic research*, *42*, 527–549. https://doi.org/10.1007/s10936-012-9231-5.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177–208. https://doi.org/10.1080/026432997381664.

Chang, M., Jones, M.N. & Johns, B.T. (2023). Comparing word frequency, semantic diversity, and semantic distinctiveness in lexical organization. *Journal of Experimental Psychology: General*, *152*(6), 1814–1823. https://doi.org/10.1037/xge0001407.

Conroy, P., Drosopoulou, C.S., Humphreys, G.F., Halai, A.D. & Lambon Ralph, M.A. (2018). Time for a quick word? The striking benefits of training speed and accuracy of word retrieval in post-stroke aphasia. *Brain*, *141*(6), 1815–1827. https://doi.org/10.1093/brain/awy087.

Dalton, S.G., Kim, H., Richardson, J. & Wright, H. (2020). A compendium of core lexicon checklists. *Seminars in Speech and Language*, *41*(1), 45–60. https://doi.org/10.1055/s-0039-3400972.

Dalton, S.G. & Richardson, J. (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. https://doi.org/10.1044/2015_AJSLP-14-0161.

Day, A. (1985). *Good dog, Carl*. New York, NY: Green Tiger Press.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838. https://doi.org/10.1037/0033-295x.104.4.801.

De Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). Reliability. *Measurement in medicine: a practical guide* (1st ed., pp. 96–146). Cambridge, UK: Cambridge University Press.

Dietz, A. & Boyle, M. (2017). Discourse measurement in aphasia: Have we reached the tipping point? *Aphasiology*, *32*(4), 459–464. https://doi.org/10.1080/02687038.2017.1398803.

Dipper, L., Marshall, J., Boyle, M., Hersh, D., Botting, N., & Cruice, M. (2021). Creating a theoretical framework to underpin discourse assessment and intervention in aphasia. *Brain Sciences*, *11*(2), 183. https://doi.org/10.3390/brainsci11020183.

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. https://doi.org/10.1044/2015_JSLHR-L-14-0249.

Fergadiotis, G., Swiderski, A. & Hula, W.D. (2019). Predicting confrontational naming difficulty. *Aphasiology*, *33*(6), 689–709. https://doi.org/10.1080/02687038.2018.1495310.

Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment*, *2*(14), 1–74. https://doi.org/10.3310/hta2140.

Fridriksson, J., den Ouden, D.-B., Hillis, A.E., Hickok, G., Rorden, C., Basilakos, A., Yourganov, G. & Bonilha, L. (2018). Anatomy of aphasia revisited. *Brain*, *141*(3), 848–862. https://doi.org/10.1093/brain/awx363.

Fromm, D., Forbes, M., Holland, A., Dalton, S.G., Richardson, J. & MacWhinney, B. (2017). Discourse characteristics in aphasia beyond the Western Aphasia Battery cutoff. *American Journal of Speech-Language Pathology*, *26*(3), 762–768. https://doi.org/10.1044/2016_AJSLP-16-0071.

Fromm, D., Greenhouse, J., Hou, K., Russell, G.A., Cai, X., Forbes, M., Holland, A. & MacWhinney, B. (2016). Automated proposition density analysis for discourse in aphasia. *Journal of Speech, Language, and Hearing Research*, *59*(5), 1123–1132. https://doi.org/10.1044/2016_JSLHR-L-15-0401.

Gahl, S. & Strand, J.F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, *89*, 162–178. https://doi.org/10.1016/j.jml.2015.12.006.

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package 'irr'. *Various coefficients of interrater reliability and agreement*, *22*, 1–32. http://bioconductor.statistik.tu-dortmund.de/cran/web/packages/irr/irr.pdf.

Gordon, J.K. (2002). Phonological neighborhood effects in aphasic speech errors: spontaneous and structured contexts. *Brain and Language*, *82*(2), 113–145. https://doi.org/10.1016/S0093-934X(02)00001-9.

Greenhalgh, K. S., & Strong, C. J. (2001). Literate language features in spoken narratives of children with typical language and children with language impairments. *Language, Speech, and Hearing Services in Schools*, *32*(2), 114–125. https://doi.org/10.1044/0161-1461(2001/010).

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010.

Henderson, A., Kim, H., Kintz, S., Frisco, N., & Wright, H. H. (2017). Working memory in aphasia: Considering discourse processing and treatment implications. *Seminars in speech and language*, 38(1), 40–51. https://doi.org/10.1055/s-0036-1597257.

Howe, T.J., Worrall, L.E. & Hickson, L.M.H. (2008). Interviews with people with aphasia: Environmental factors that influence their community participation. *Aphasiology*, *22*(10), 1092–1120. https://doi.org/10.1080/02687030701640941.

Hudspeth, S.G., Dillow, E. & Richardson, J.D. (2013). Core lexicon analysis: efficient assessment of narrative discourse in persons with aphasia. Technical research session presented at Annual Convention of the American Speech-Language-Hearing Association. Chicago, IL.

Johns, B.T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review*, *128*(3), 525–557. https://doi.org/10.1037/rev0000265.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, *66*, 115–124. http://doi.org/10.1037/a0026727.

Jusczyk, P.W., Luce, P.A. & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645. https://doi.org/10.1006/jmla.1994.1030.

Kagan, A., Simmons-Mackie, N., Rowland, A., Huijbregts, M., Shumway, E., McEwen, S., Threats, T. & Sharp, S. (2008). Counting what counts: A framework for capturing real-life outcomes of aphasia intervention. *Aphasiology*, *22*(3), 258–280. https://doi.org/0.1080/02687030701282595.

Kasselimis, D.S., Simos, P.G., Peppas, C., Evdokimidis, I. & Potagas, C. (2017). The unbridged gap between clinical diagnosis and contemporary research on aphasia: A short discussion on the validity and clinical utility of taxonomic categories. *Brain and Language*, *164*, 63–67. https://doi.org/10.1016/j.bandl.2016.10.005.

Kendall, D. L., Rosenbek, J. C., Heilman, K. M., Conway, T., Klenberg, K., Rothi, L. J. G., & Nadeau, S. E. (2008). Phoneme-based rehabilitation of anomia in aphasia. *Brain and language*, *105*(1), 1–17. https://doi.org/10.1016/j.bandl.2007.11.007.

Kim, H., Berube, S., & Hillis, A. E. (2023). Core lexicon in aphasia: A longitudinal study. *Aphasiology*, *37*(10), 1679–1691. https://doi.org/10.1080/02687038.2022.2121598.

Kim, H., Schoemann, A.M. & Wright, H.H. (2022). Quality of measurement in core lexicon measures. *Journal of Speech, Language, and Hearing Research*, *65*(8), 2919–2930. https://doi.org/10.1044/2022_JSLHR-20-00722.

Kim, H. & Wright, H.H. (2020a). A tutorial on core lexicon: Development, use, and application. *Seminars in Speech and Language*, *41*(1), 20–31. https://doi.org/10.1055/s-0039-3400973.

Kim, H., & Wright, H. H. (2020b). 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. https://doi.org/10.1044/2019_AJSLP-19-0063.

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. https://doi.org/10.1111/1460-6984.12432.

Koo, T.K. & Li, M.Y. (2016). A guideline for selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kreisler, A., Godefroy, O., Delmaire, C., Debachy, B., Leclercq, M., Pruvo, J.-P. & Leys, D. (2000). The anatomy of aphasia revisited. *Neurology*, *54*(5), 1117–1123. https://doi.org/10.1212/WNL.54.5.1117.

Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990. https://doi.org/10.3758/s13428-012-0210-4.

Lampe, L.F., Hameau, S. & Nickels, L. (2022). Semantic variables both help and hinder word production: Behavioral evidence from picture naming. *Journal of experimental psychology: Learning, memory, and cognition*, *48* (1), 72–97. https://doi.org/10.1037/xlm0001050.

Levelt, W.J.M., Roelofs, A. & Meyers, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776.

Liljequist, D., Elfving, B. & Skavberg Roaldsen, K. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLoS ONE*, *14*(7), e0219854. https://doi.org/10.1371/journal.pone.0219854.

Linnk, A., Bastiaanse, R. & Höhle, B. (2016). Discourse production in aphasia: A current review of theoretical and methodological challenges. *Aphasiology*, *30*(7), 765–800. https://doi.org/10.1080/02687038.2015.1113489.

Linnik, A., Bastiaanse, R., Stede, M., & Khudyakova, M. (2022). Linguistic mechanisms of coherence in aphasic and non-aphasic discourse. *Aphasiology*, *36*(2), 123–146. https://doi.org/10.1080/02687038.2020.1852527.

Loring, D. W., Lah, J. J., & Goldstein, F. C. (2023). Telehealth equivalence of the Montreal cognitive assessment (MoCA): Results from the Emory healthy brain study (EHBS). *Journal of the American Geriatrics Society*, *71*(6), 1931–1936. https://doi.org/10.1111/jgs.18271.

Lowell, S., Beeson, P. M., & Holland, A. L. (1995). The efficacy of a semantic cueing procedure on naming performance of adults with aphasia. *American Journal of Speech-Language Pathology*, *4*(4), 109–114. https://doi.org/10.1044/1058-0360.0404.109.

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*, 439–463. https://doi.org/10.1007/s10936-005-6203-z.

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

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30.

McRae, K., Cree, G.S., Seidenberg, M.S. & McNorgan, C. **(**2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. https://doi.org/10.3758/BF03192726.

Middleton, E.L. & Schwartz, M.F. (2010). Density pervades: An analysis of phonological neighbourhood density effects in aphasic speakers with different types of naming impairment. *Cognitive Neuropsychology*, *27*(5), 401–427. https://doi.org/10.1080/02643294.2011.570325.

Milani, S. A., Marsiske, M., Cottler, L. B., Chen, X., & Striley, C. W. (2018). Optimal cutoffs for the Montreal Cognitive Assessment vary by race and ethnicity. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *10*, 773–781. https://doi.org/10.1016%2Fj.dadm.2018.09.003.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M. & De Vet, H.C.W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*, 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.

Mullen, L.A., Benoit, K., Keyes, O., Selivanov, D. & Arnold, J. (2018). Fast, consistent tokenization of natural language text. *Journal of Open Source Software*, *3*, 655. https://doi.org/10.21105/joss.00655.

Naeser, M. A. & Hayward, R. W. (1978). Lesion localization in aphasia with cranial computed tomography and the Boston Diagnostic Aphasia Exam. *Neurology*, *28*(6), 545–545. https://doi.org/10.1212/wnl.28.6.545.

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L. & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, *53*, 695-699. https://doi.org/10.1111/j.1532-5415.2005.53221.x.

Newman, R.S. & German, D.J. (2005). Life span effects of lexical factors on oral naming. *Language and Speech*, *48*(2), 123–156. https://doi.org/10.1177/00238309050480020101.

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, Language, and Hearing Research*, *36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338.

Nickels, L. (2002). Therapy for naming disorders: Revising, revising, and reviewing. *Aphasiology*, *16*(10/11), 935–979. https://doi.org/10.1080/02687030244000563.

Pavelko, S. L., Price, L. R., & Owens Jr, R. E. (2020). Revisiting reliability: Using Sampling Utterances and Grammatical Analysis Revised (SUGAR) to compare 25-and 50-utterance language samples. *Language, Speech, and Hearing Services in Schools*, *51*(3), 778–794. https://doi.org/10.1044/2020_LSHSS-19-00026.

Pritchard, M., Hilari, K., Cocks, N. & Dipper, L. (2018). Psychometric properties of discourse measures in aphasia: Acceptability, reliability, and validity. *International Journal of Language & Communication Disorders*, *53*(6), 1057–1169. https://doi.org/10.1111/1460-6984.12420.

Rabovsky, M., Schad, D.J. & Rahman, R.A. (2016). Language production is facilitated by semantic richness but inhibited by semantic density: Evidence from picture naming. *Cognition*, *146*, 240–244. https://doi-org.pitt.idm.oclc.org/10.1016/j.cognition.2015.09.016.

Ratcliffe, L. N., McDonald, T., Robinson, B., Sass, J. R., Loring, D. W., & Hewitt, K. C. (2023). Classification statistics of the Montreal Cognitive Assessment (MoCA): Are we interpreting the MoCA correctly? *The Clinical Neuropsychologist*, *37*(3), 562–576. https://doi.org/10.1080%2F13854046.2022.2086487.

Schmid, H. (1997). Probabilistic part-of-speech tagging using decision trees. In D. B. Jones & H. Somers (Ed.). *New methods in language processing* (1st ed., pp. 154–164). London, UK: Routledge.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Smith, K.G. & Bhutada, A.M. (2021). Detailed vision screening results from a cohort of individuals with aphasia. *Aphasiology*, *35*(2), 186–199. https://doi.org/10.1080/02687038.2019.1702918.

Spencer, T. D., Tolentino, T. J., & Foster, M. E. (2023). Impact of discourse type and elicitation task on language sampling outcomes. *American Journal of Speech-Language Pathology*, *32*(6), 2827–2845. https://doi.org/10.1044/2023_AJSLP-22-00365.

Stark, B. C., Bryant, L., Themistocleous, C., Den Ouden, D. B., & Roberts, A. C. (2023a). Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology*, *37*(5), 761-784. https://doi.org/10.1080/02687038.2022.2039372.

Stark, B.C., Alexandar, J.M., Hittson, A., Doub, A., Igleheart, M., Streander, T. & Jewell, E. (2023b). Test-retest reliability of microlinguistic information derived from spoken discourse in persons with chronic aphasia. *Journal of Speech, Language, and Hearing Research*, *66*(7), 2316–2345. https://doi.org/10.1044/2023_JSLHR-22-00266.

Stark, B. C., Dutta, M., Murray, L. L., Bryant, L., Fromm, D., MacWhinney, B., Ramage, A.E., Roberts, A., den Ouden, D.B., Brock, K., McKinney-Bock, K., Paek, E.J., Harmon, T.G., Yoon, S.O., Themistocleous, C., Yoo, H., Aveni, K., Gutierrez, S. & Sharma, S. (2021). Standardizing assessment of spoken discourse in aphasia: A working group with deliverables. *American Journal of Speech-Language Pathology, 30*(1S), 491-502. https://doi.org/10.1044/2020_AJSLP-19-00093.

Stark, B.C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology*, *28*(3), 1067–1083. https://doi.org/10.1044/2019_AJSLP-18-0265.

Storkel, H., Armbrüster, J. & Hogan, T.P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192. https://doi.org/10.1044/1092-4388(2006/085).

Stratford, P.W. (2004). Getting more from the literature: Estimating the standard error of measurement from reliability studies. *Physiotherapy Canada*, 56, 27–30.

Székely, A., D'Amico, S., Devescovi, A., Federmeier, K., Herron, D., Iyer, G., Jacobsen, T. & Bates, E. (2003). Timed picture naming: Extended norms and validation against previous studies. *Behavior Research Methods, Instruments, & Computers*, *35*, 621–633. https://doi.org/10.3758/BF03195542.

Tan, S. (2007). *The arrival*. New York, NY: Arthur A. Levine Books.

Vaden, K.I., Halpin, H.R. & Hickok, G.S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file: https://www.iphod.com/download/IPhODv2.0_REALS.zip]. Available from http://www.iphod.com.

Vitevitch, M.S., Armbrüster, J. & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 514–529. https://doi.org/10.1037/0278-7393.30.2.514.

Vitevitch, M.S. & Sommers, M.S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition, 31*, 491–504. https://doi.org/10.3758/BF03196091.

Vitevitch, M.S. & Luce, P.A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*, 374–408. https://doi.org/10.1006/jmla.1998.2618.

Weir, J.P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J. Strength Cond. Res.*, *19*(1), 231–240. https://doi.org/10.1519/15184.1.

World Health Organization International Classification of Functioning, Disability and Health (ICF). Geneva, Switzerland: World Health Organization; 2001.

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)

Ylvisaker, M., & Szekeres, S. (1985, November). Cognitive-language intervention with brain-injured adolescents and adults. Mini-seminar presented at the annual convention of the Illinois Speech-Language-Hearing Association.