Article accepté dans le Journal of Speech, Language, and Hearing Research
DOI: 10.1044/2025_JSLHR-24-00713

Core lexicon in Laurentian French speakers without brain injury: Development, validation, and reliability

Amélie Brisebois[1,2], Simona Maria Brambati[3,4,5], Éva Marois[2], and Karine Marcotte[1,2]

[1] École d'orthophonie et d'audiologie, Faculté de médecine, Université de Montréal, Montréal, Québec, Canada.

[2] Centre de recherche du Centre intégré universitaire de santé et de services sociaux du Nord-de-l'Île-de-Montréal, Montréal, Québec, Canada.

[3] Faculté de médecine, Université de Montréal, Montréal, Québec, Canada.

[4] Centre de recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, Québec, Canada.

[5] Département de psychologie, Faculté des arts et des sciences, Université de Montréal, Montréal, Québec, Canada

***Correspondence:**

Amélie Brisebois, MHSc

Address: Centre de recherche du Centre intégré universitaire de santé et de services sociaux du Nord-de-l'Île-de-Montréal, 5400 Gouin Ouest, Montréal, Québec, Canada, H4J 1C5

Phone number: 514-338-2222 extension 7710, Fax number: 514-340-2115

E-mail: amelie.brisebois@umontreal.ca

Disclosure: The authors have declared that no competing financial or nonfinancial interests existed at the time of publication

Abstract

Purpose: Lexical performance in discourse is of considerable interest in acquired communication disorders. The transcription-free core lexicon measure evaluates the most typical words a person uses during communication. This study aimed (1) to develop core lexicon lists in Laurentian French speakers without brain injury and (2) to assess their psychometric properties.

Method: Spoken discourse was elicited using the picture description task from the Western Aphasia Battery-Revised (WAB-R; Kertesz, 2006) and the Cinderella Story Telling task (CST). Participants were Laurentian French speakers from Quebec, aged 50–79, without brain injury (PWBI). Sixty-six completed the WAB-R task and 48 completed the CST task. Core noun and verb lists were created using the CLAN program, including words produced by at least 50% of the sample. Two raters scored all audio samples. Intra- and inter-rater reliability and long-term test–retest reliability were calculated. Construct validity was examined through correlations with micro- and macrostructural discourse measures.

Results: Four core lexicon lists were generated. For the WAB-R, 19 nouns and 5 verbs were identified; for the CST, 19 nouns and 16 verbs. Intra-rater reliability was excellent across variables, and inter-rater reliability was excellent for all core noun lists and CST core verbs, and good for WAB-R core verbs. Long-term test–retest reliability ranged from poor to moderate across measures. Core lexicon scores were significantly and positively correlated with 12 macrostructural and 9 microstructural variables.

Conclusions: This study supports the rater reliability and construct validity of core

lexicon measures in Laurentian French across two discourse tasks. It also provides the first long-term test–retest reliability data for core lexicon scoring, offering insights that guide its clinical and research applications.

Keywords: Discourse analysis, core lexicon, test-retest reliability

Introduction

Discourse analysis in people with acquired language difficulties provides valuable information about language function and impairment and insight into everyday communication (Armstrong, 2000). In the current study, we refer to discourse as language beyond the sentence level (Armstrong, 2000). In clinical settings, discourse assessment is essential to a comprehensive communication assessment (Bryant et al., 2016), but clinicians are reluctant to perform discourse analysis because it is generally time-consuming. Another fundamental challenge regarding the clinical implementation of discourse evaluation is the lack of psychometric documentation (Boyle, 2020). The quality of discourse measures depends on many factors, including psychometric features such as reliability and validity (Pritchard et al., 2017).

Measures in discourse

In acquired neurogenic communication disorders, discourse measures have been particularly useful in identifying very mild language impairment (Fromm et al., 2017) and distinguishing between healthy aging persons without brain injury (PWBI) and persons with mild cognitive impairment (e.g., Forbes-McKay & Venneri, 2005; Kim et al., 2022; Mueller et al., 2018; Taler et al., 2021). Discourse samples can also guide the diagnostic classification of primary progressive aphasia (Wilson et al., 2010), help document cognitive changes amongst persons with cognitive impairment (Antonsson et al., 2021), and inform about the language of people with Alzheimer's disease (e.g., Kim & Lee, 2023; Slegers et al., 2018).

On the theoretical level, discourse production is divided into three distinct stages: (a) conceptual preparation, (b) linguistic formulation, and (c) articulation and monitoring of

the verbal message (Frederiksen & Stemmer, 1993). The conceptual preparation stage also refers to what others call the macrostructure, whereas the linguistic formulation relates to the microstructure of a text. Macrostructural measures concern the discourse-level organization features such as informativeness, coherence, and cohesion, whereas microstructural measures represent within-sentence features and depict discourse's lexical and grammatical components.

Core lexicon in English and other languages

Core lexicon items are key lexical elements used during a discourse task that make a language sample relevant and coherent. They are classified as microstructural variables, as they represent the lexical component of discourse.

MacWhinney et al. (2010) were the first to explore core lexicon by comparing the top ten most frequently produced nouns and verbs in the Cinderella Story Telling (CST) task between 25 PWBI and 24 individuals with aphasia (PWA). Rather than creating a single list for scoring, they compared the rank order of word frequencies across groups, reporting that PWA used fewer lexical items overall and tended to rely on more general and lighter verbs. Fromm et al. (2013) applied a similar ranking approach using the procedural Sandwich task in a much larger sample (144 PWBI, 141 PWA), finding that while the types of words produced were similar across groups, their frequency rankings differed.

Subsequent studies have employed more systematic methods for list construction and scoring. Dalton and Richardson (2015) created core lexicon lists by identifying lemmas (regardless of word class) produced by ≥ 50% of 92 PWBI across five discourse tasks

from the AphasiaBank protocol. Lemmas are the base or dictionary forms of a word under which all its inflected variants are grouped (Matthews, 2007). Kim et al. (2019) constructed lists of the 25 most frequent lemmas for each word class—nouns, verbs, adjectives, and adverbs—based on samples from 470 PWBI retelling two wordless picture books (Good Dog Carl and Picnic). These lists were used to score data from 11 PWA, revealing age-related effects and a positive correlation between verb production and Aphasia Quotient (WAB-R; Kertesz, 2006) scores. Fluent aphasia was associated with more verb production than non-fluent aphasia. In a follow-up, Kim et al. (2020) found that PWA produced fewer function words than controls. Kim et al. (2022) used the original Cookie Theft picture description task to study 19 PWA longitudinally and found core lexicon scores improved from the acute to chronic stages of recovery. Dalton et al. (2024) also created task-specific lists from 45 and 50 PWBI for the original and modern versions of the Cookie Theft task, using a threshold of ≥ 50% occurrence. In these later studies, core lexicon scoring involved assigning one point per target word used by a clinical participant, based on control-derived lists.

More recently, core lexicon methods have been extended to populations with cognitive impairment. Kintz et al. (2024) conducted a preliminary study on 12 individuals with suspected Alzheimer's disease, showing that lower core lexicon scores were associated with greater dementia severity and poorer language performance. In a larger study, Fromm et al. (2024) reported that 122 individuals with Alzheimer's disease and 15 with mild cognitive impairment produced significantly fewer core lexicon items than PWBI when describing the Cookie Theft picture.

Together, these studies illustrate the value of core lexicon scoring in distinguishing

between PWBI and individuals with aphasia or cognitive decline, and in capturing meaningful variation within aphasia subtypes. They also highlight the importance of transparency and consistency in list development and scoring procedures to support future cross-study comparisons and clinical applications.

Core lexicon checklists have been developed in Mandarin for seven discourse tasks, including picture descriptions, story narratives, and a procedural task (Chen & Chang, 2024; Jiang et al., 2023). Jiang et al. created checklists of the 25 most frequent nouns and verbs from 88 PWBI and showed that 12 PWA produced significantly fewer core items across tasks. Chen and Chang (2024) selected the 30 words with the widest distribution across the normative sample of 43 PWBI for each of the seven tasks to construct core lexicon checklists. Core lexicon scores were significantly correlated with lexical diversity and discourse informativeness, supporting construct validity. These studies also revealed linguistic distinctions in Mandarin, such as the frequent inclusion of the function word 'le' and aspect markers—items not typically found in English core lexicons.

These findings emphasize the importance of developing language-specific core lexicon lists. The approach is promising for clinical use due to its quick and intuitive, transcription-less scoring (Dalton et al., 2020). Building on evidence that transcription-free measures can successfully distinguish between PWBI and adults with mild cognitive impairment (Kim et al., 2022), real-time scoring would be a key advantage for clinical application. This article focuses on developing the clinically accessible core lexicon measure in Laurentian (Quebec) French, including psychometric validation: intra- and inter-rater reliability, construct validity, and long-term test-retest reliability.

Psychometric quality of core lexicon

Psychometric characteristics of core lexicon measures have been explored in recent studies. Concerning intra- and inter-rater reliability, the core lexicon measure is expected to be excellent since scoring relies on the presence of a closed list of lexical items. Inter-rater reliability was reported on ten language samples of PWA for two story telling tasks (Kim & Wright, 2020). The stimuli were the wordless picture books of *Good Dog Carl* and *Picnic*. Scoring was performed by four raters who listened to each audio file twice for each core lexicon list. This procedure was chosen to approximate the typical clinical time to complete an assessment. Intra-class correlations were above .90 for all core lexicon lists.

Two main approaches exist: frequency-based (counting all occurrences of each lemma) and percentage-based (identifying lemmas produced by a set proportion of participants, e.g., 50% or 75%). While the optimal method remains inconclusive (Chen & Chang, 2024), both yield strong evidence of construct validity (Kim et al., 2022). Several studies support this. Alyahya et al. (2021) showed very high correlations between lexicon landscapes and correct information units (CIU) in both PWBI and PWA. Similarly, core lexicon has demonstrated strong associations with widely used discourse metrics such as Main Concepts (MC) and CIU (Dalton et al., 2015). Kim and Wright (2020) also reported significant correlations between core lexicon measures and discourse variables including syntactic complexity, lexical diversity, coherence, thematic units, and information units. Together, these findings indicate that core lexicon measures validly and efficiently assess lexical retrieval by capturing the key lexical elements that support the relevance and coherence of discourse.

Aims of the study

The present study had two main objectives:

(1) To develop core lexicon lists of nouns and verbs for the Cinderella Story Telling (CST) task and the Picnic picture description from the Western Aphasia Battery-Revised (WAB-R) using samples from Laurentian French speakers without brain injury (PWBI); and

(2) To evaluate the psychometric properties of the core lexicon measure, including intra-rater, inter-rater, and long-term test–retest reliability, construct validity, and minimal detectable change at the 90% confidence level (MDC90).

Both discourse tasks were selected to capture distinct elicitation contexts—a story telling (CST) and a picture description (WAB-R). A recent study by Schnur and Wang (2024) found that these tasks yield divergent discourse profiles: CST elicited more lexically diverse, structurally complex, and syntactically accurate speech than WAB-R, which tended to prompt shorter, list-like utterances. These differences are attributed to the inherent cognitive and linguistic demands of each task, with CST offering less visual support and requiring greater spontaneous language generation. Including both tasks thus allows us to sample a broader range of discourse behaviors across contexts. In line with previous approaches (e.g., MacWhinney et al., 2010; Alyahya et al., 2021), we generated separate noun and verb core lexicon lists, a distinction particularly relevant in French due to its complex verb morphology. Core lexicon efficiency variables—nouns and verbs per minute—were included to index informativeness relative to speaking time, responding to recent calls to integrate time-sensitive discourse metrics into clinical research and

practice (Dalton et al., 2020). Finally, minimal detectable change at the 90% confidence level (MDC90) was calculated to support the interpretation of individual-level change in test–retest procedures.

Methods

The manuscript reports all necessary and recommended standards for reporting spoken discourse. Supplementary Material 1 provides the best practice guidelines checklist from Stark et al. (2022).

Participants

All participants were recruited as control PWBI in larger projects approved by the ethics committee at Centre de recherche du Centre intégré universitaire de santé et de services sociaux du Nord-de-l'Île-de-Montréal (CIUSSS-NÎM). The discourse samples consisted of subsets of PWBI from previously published studies (CST: Brisebois et al., 2023; WAB-R: Marcotte et al., 2022, 2024). For the WAB-R task, 66 PWBI were recruited as controls: 18 for a study on longitudinal changes in post-stroke aphasia (CIUSSS-NÎM # MP-32-2018-1478) and 48 during the COVID-19 pandemic for a study on longitudinal changes in spoken discourse (CIUSSS-NÎM # 2020-1900). This group included 38 females and 28 males, with a mean age of 64.5 ± 7.2 years and average education of 16.1 ± 2.9 years. The 48 CST participants, recruited under the same ethics protocol (# 2020-1900), included 30 females and 18 males (mean age: 64.3 ± 6.6 years; mean education: 16.3 ± 2.8 years). Written informed consent was obtained from all participants. Inclusion criteria were: (1) age 50 or older and (2) Laurentian (Quebec) French as their primary language of use. Exclusion criteria were: (1) severe mental illness, (2) acquired or developmental language impairment, (3) neurological or neurocognitive disorders, (4)

traumatic brain injury, (5) self-reported cognitive complaints, and (6) uncorrected vision or hearing issues. Exclusion criteria were assessed via self-report questionnaires. All participants were Caucasian.

Long-term test–retest procedure

All 66 WAB-R participants completed a retest session, on average 253.4 ± 67.5 days after the initial assessment. For the CST task, 45 out of 48 participants completed a retest, on average 239.0 ± 56.9 days later. All participants self-reported no health changes between the sessions, including cognitive or language changes. We recognize that test–retest reliability studies often use short intervals (~ 2 weeks) to reduce the influence of external factors such as aging or learning. However, our decision to use a longer interval was deliberate and grounded in considerations of ecological validity. In aging populations or in contexts involving neurodegenerative disease, reassessments typically occur over several months (e.g., Alioto et al., 2017). Shorter intervals may inflate reliability due to practice effects, especially in unimpaired participants (Calamia et al., 2012). Moreover, test–retest stability can vary substantially between control and clinical groups (Stark et al., 2023), limiting the generalizability of reliability metrics from unimpaired samples to individuals with aphasia.

Data collection and transcription

Data collection and transcription procedures for both tasks of the Cinderella Story Telling (CST) and the picnic picture description (WAB-R; Kertesz, 2006) are fully reported in previously published studies (CST: Brisebois et al., 2023; WAB-R: Marcotte et al., 2024). Participants were assessed under the same conditions for test and retest. Briefly, the video/audios of each discourse sample were imported and transcribed in ELAN

(Sloetjes & Wittenburg, 2008) using CHAT conventions. Complete orthographic transcriptions were conducted, and the transcription was verbatim. The CHAT manual (MacWhinney, 2000) was used for utterance segmentation, transcription, and scoring, with additional guidance for French speakers (Colin & Le Meur, 2016). For the WAB-R task, the mean duration of the samples was 84.58 seconds (range: 26-202; SD = 40.48) and included a mean of 229.47 words (range: 72-658; SD = 113.91). For the CST task, the mean duration of the samples was 183.98 seconds (range: 21-423; SD = 79.81) and included a mean of 450.48 words (range: 77-1150; SD = 200.19). These results were calculated at the first session (test).

Lemma extraction

Once the transcription of the first session was completed, the morphological and grammatical information was coded using the CLAN (MacWhinney, 2000) program *mor*, which tags morphemes and lemmas under each transcript utterance. Subsequently, all lemmas were extracted for each sample using CLAN. Lemmas were visually inspected, and inflections (verbal inflections, gender, or plural markers) of the same lemma were merged. For the WAB-R task, the 66 participants produced 15,145 words (tokens) and 616 unique lemmas (456 nouns and 160 verbs). For the CST task, the 48 participants produced 21,623 words (tokens) and 808 unique lemmas (546 nouns and 262 verbs).

Core lexicon lists

Following prior studies (Alyahya et al., 2021; Dalton et al., 2024; Dalton & Richardson, 2015), we created core lexicon lists based on the percentage of participants who produced each lemma. We generated four lists: core nouns and core verbs for each discourse task, including only words produced by more than 50% of PWBI participants.

Core lexicon Scoring

Like Kim and Wright (2020), the audio recordings were used to manually score each sample using a Microsoft Excel spreadsheet (provided in Supplementary Materials 2). Both raters (A.B. and E.M.) trained on approximately 10 samples separately and discussed potential issues before performing the final scoring of the first assessment (test). Since the second round of scoring for intra-rater agreement occurred approximately 12 weeks apart, no refresher session was needed. The same rater scored test and retest of the same participant. To approximate clinical scoring conditions, two raters (A.B. and E.M.) listened twice to each audio sample for each list of core lexicon items. Participants received one point for each core lexicon item produced—excluding synonyms and including inflections—per item.

Core lexicon variables

The core lexicon measures included raw scores for both nouns and verbs as well as time-based efficiency measures: core lexicon nouns per minute and verbs per minute (Dalton et al., 2020). These variables reflect the efficiency with which key lexical items are produced during discourse. Time-based efficiency measures, such as CIUs/min, are well-established and clinically relevant, as they capture speaker effort and listener processing demands while requiring minimal clinician effort. Extending this approach to core lexicon items offers a promising, though still exploratory, method for assessing lexical retrieval in discourse, particularly when sample durations are sufficient to support stable estimates.

Dependent variables for Construct Validity

Macrostructural variables

For the CST task, we used the MC scoring system adapted to Laurentian French (Brisebois et al., 2023). The current study used total MC score (MC_total) and derived efficiency measure (MC_total per minute; MC_min). For the WAB-R picnic picture description, the Thematic Units (TU) variable and its derived efficiency measure (TU per minute; TU_min) were used (Brisebois et al., 2020). Definitions of these variables appear in Table 1.

Table 1. Definition of the macro- and micro-structural variables.

	Measure	Definition	Language dimension
Macrostructural measures	MC_total	The total score of the Main concepts in the Cinderella Story Retell task	Main Concept
	MC_min	The number of Main concepts per minute in the Cinderella Story Retell task	Main Concept efficiency
	TU	The total of Thematic Units in the WAB-R task	Macrostructural informativeness
	TU_min	The number of Thematic Units per minute in the WAB-R task	Mactrostructural informativeness efficiency
Microstructural measures	CIU	Correct Information Units	Lexical informativeness
	Moving Average Token-Type Ratio (MATTR)	Average of estimated Token-Type Ratios for successive nonoverlapping successive windows of fixed length	Lexical diversity
	Verbs per utterance	Average number of verbs (verbs, copulas, auxiliaries followed by past or present participles) per utterance.	Syntactic complexity

Note. Microstructural data was derived from the CLAN software (MacWhinney et al., 2010), including CIU which were extracted similarly to Deng et al., 2024.

Microstructural variables

The selection of microstructural variables was inspired by other concurrent validity

investigations of core lexicon (Kim & Wright, 2020; Alyahya et al., 2021). These variables are described in Table 3 and include Correct information units (CIU; Nicholas & Brookshire, 1993) and derived efficiency measures of CIU per minute (CIU_min), the Moving Average Token-Type Ratio (MATTR; Covington, 2007), and the number of verbs per utterance.

Data analysis

Intra- and inter-rater reliability

To determine intra- and inter-rater reliability in core lexicon scoring, 23 samples per rater (representing approximately 20% of the samples; a total of 13 samples for the WAB-R task and 10 samples for the CST task) were randomly selected for each of the two raters. For intra-rater reliability, raters scored core lexicon items twice, approximately 12 weeks apart (MEAN = 87.5; SD = 3.0 days) between June and September 2024. As for inter-rater reliability, E.M. initially scored samples by A.B. and vice versa.

Interrater reliability for the dependent variables of the construct validity have been thoroughly documented in previous studies, including MC (Brisebois et al., 2023) and Thematic Units (Marcotte et al., 2024).

Statistical analysis

All statistical analyses were performed using SPSS® v29.0. and the significance level was set at $p < .05$.

Data distribution was analysed using Kolmogorov-Smirnov test for all variables (MC_total, MC_min, TU, TU_min, CIU, CIU_min, MATTR, number of verbs per utterance, core lexicon verbs, core lexicon nouns, core lexicon verbs per minute and core lexicon nouns per minute) for each session. More than 80% of the variables were non-

normally distributed and non-parametric tests were used throughout.

Following the guidelines of Koo and Li (2016) to select the appropriate ICC, intra and inter-rater coding reliability were evaluated using two-way mixed ICC with absolute agreement.

Construct validity analyses were conducted using the first assessment dataset (test). Following the approach of Kim and Wright (2020), validity was assessed using Spearman's rho correlations between core lexicon scores and a range of micro- and macrostructural discourse variables. Specifically, we examined relationships with total units (TU),CIU, andMC, as these are established markers of lexical informativeness and structural content in discourse. Including these comparisons allowed us to explore the extent to which core lexicon measures (including efficiency scores) align with or diverge from other discourse indicators, thereby contributing to their clinical and theoretical interpretability.

For long-term test-retest, reliability was assessed with two-way mixed ICC absolute agreement. Agreement was tested using the Wilcoxon signed-rank test to evaluate if there was a statistically significant difference between test and retest. We also measured the strength of association using Spearman's rho to assess similarity between test and retest. The significance level was set at $p < .05$.

As core lexicon lists could help detect subclinical language or cognitive deficits, we also provided minimal detectable change (MDC) for each core lexicon list. Given the variance from the test-retest results, MDC at a 90% confidence interval (CI) (MDC90) was computed to assess the approximate change needed to be associated with clinical change.

MDC90 includes the standard error of measurement (SEM), computed with the following formula: SEM = SD √ 1 - *r*, where SD is the standard deviation for the obtained score distribution and *r* is the correlation coefficient (i.e., ICC). The formula to calculate MDC90 is MDC90 = SEM ∗ 1.65 ∗ √ (2).

Results

Development of the core lexicon verbs and nouns lists

Analysis of core lexicon

Tables 2 and 3 provide a list of each verb and noun lemma produced by more than 50% of the sampling cohorts along with its frequency and the number of participants who produced the lexeme for each task.

Table 2. Frequency, number and percentage of participants who produced each core noun and core verb for the Cinderella Story Telling Task.

Core nouns	Frequency	n (max=48)	%
Prince [prince]	188	48	100.00%
Minuit [midnight]	105	46	95.83%
Cendrillon [Cinderella]	307	45	93.75%
Fille [girl]	238	45	93.75%
Fée [fairy]	78	44	91.67%
Robe [dress]	136	42	87.50%
Bal [ball]	210	41	85.42%
Soulier [shoe]	154	41	85.42%
Carrosse [carriage]	75	38	79.17%
Enfant [child]	47	34	70.83%
Maison [house]	78	33	68.75%
Citrouille [pumpkin]	49	33	68.75%
Verre [glass]	71	31	64.58%
Soeur [sister]	74	28	58.33%
Château [castle]	57	26	54.17%
Souris [mouse]	47	25	52.08%
Histoire [story]	40	25	52.08%
Mère [mother]	37	25	52.08%
Belle-mère [stepmother]	75	24	50.00%

Core nouns	Frequency	n (max=48)	%
Core verbs			
Être [be]	694	48	100.00%
Avoir [have]	645	48	100.00%
Aller [go]	282	46	95.83%
Faire [do]	139	40	83.33%
Vouloir [want]	74	37	77.08%
Marier [marry]	54	37	77.08%
Essayer [try]	60	34	70.83%
Trouver [find]	66	32	66.67%
Falloir [need]	50	31	64.58%
Voir [see]	53	30	62.50%
Devoir [must]	50	29	60.42%
Pouvoir [can]	67	28	58.33%
Savoir [know]	40	27	56.25%
Arriver [arrive]	50	24	50.00%
Retrouver [find]	46	24	50.00%
Perdre [lose]	24	24	50.00%

Table 3. Frequency counts. number and percentage of participants who produced each Core Noun and Core verb for the picture description task of the WAB-R.

Core Nouns	Frequency	n (max=66)	%
Voilier [sailing ship]	92	66	100.00%
Cerf-volant [kite]	90	65	98.48%
Chien [dog]	78	65	98.48%
Château [castle]	74	64	96.97%
Sable [sand]	79	61	92.42%
Fille [girl]	70	58	87.88%
Garçon [boy]	80	57	86.36%
Pique-nique [picnic]	88	56	84.85%
Voiture [car]	63	55	83.33%
Maison [house]	135	53	80.30%
Radio [radio]	56	52	78.79%
Poisson [fish]	58	51	77.27%
Bord [shore]	86	46	69.70%
Arbre [tree]	70	44	66.67%
Quai [dock]	53	41	62.12%
Drapeau [flag]	44	40	60.61%
Eau [water]	71	37	56.06%

Core Nouns	Frequency	n (max=66)	%
Monsieur [man]	67	35	53.03%
Lac [lake]	64	33	50.00%
Core Verbs			
Avoir [to have]	606	66	100.00%
Être [to be]	606	66	100.00%
Faire [to do]	77	47	71.21%
Voir [to see]	134	42	63.64%
Verser [to pour]	37	35	53.03%

Intra and inter-rater reliability

Koo and Li (2016) interpretation guidelines were used for all ICCs (intra- and inter-rater reliability): below .50 = poor; between .50 and .75 = moderate; between .75 and .90 = good; and above .90 = excellent. Intra- and inter-rater reliability (IRR) were calculated for each discourse task's lemma list (nouns and verbs). Intra-rater reliability ICCs were excellent for all lists and the two raters, except for the core nouns list of the WAB-R that was good for one rater (ICC = .892). As for inter-rater reliability, the core nouns and verbs for CST and the core verbs for the WAB-R met the threshold of excellent reliability, ICC ≥ .90, and the core nouns for the WAB-R were good (ICC = .878). ICC and Standard Error Measurement for each list are reported in Table 4.

Table 4. Inter-rater reliability results for all Core Lexicon lists.

Discourse task	Measure	Nouns	Verbs
Cinderella Story Retell	ICC	.989	.943
	SEM	0.31	0.61
Picnic picture description of the WAB-R	ICC	.878	.922
	SEM	1.93	1.55

All ICCs are positive and significant (p < .001). SEM = standard error of measurement.

Construct Validity Analyses

For the CST task, core nouns significantly correlated with MC_total and CIU. Core nouns per minute significantly correlated with MC_total, the MC_min, and CIU. Core verbs significantly correlated with MC_total, CIU, MATTR and Verbs_Utt. Core verbs per minute significantly correlated with the MC_total, the MC_min, and CIU. These results appear in Table 5.

Table 5. Spearman correlation results for the Cinderella story retell task.

	MC_total	MC_min	CIU	MATTR	Verbs_Utt
Core nouns	.586**	-.096	.629**	.044	.064
Core verbs	.609**	-.103	.654**	-292*	.425**
Core nouns per minute	-.518**	.427**	-.699**	.204	-.218
Core verbs per minute	-.573**	.389**	-.767**	.013	-.005

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

For the WAB-R task, Spearman analyses revealed that core nouns significantly correlated with TU, TU_min, and CIU. Core nouns per minute significantly correlated with TU_min and CIU. Core verbs also significantly correlated with TU, TU_min, and CIU. Core verbs per minute significantly correlated with the TU_min and CIU. These results appear in Table 6.

Table 6. Spearman correlation results for the Picnic picture description of the WAB-R task.

	TU	TU_min	CIU	MATTR	Verbs_Utt
Core nouns	.649**	-.494*	.657**	.104	.083
Core verbs	.428**	-.191**	.344**	-.075	.083
Core nouns per minute	-.090	.950**	-.771**	-.058	-.158
Core verbs per minute	-.179	.940**	-.795**	-.054	-.119

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

Long-term test-retest reliability analyses

The descriptive statistics of each core lexicon variable (data distribution, means, standard deviations, ranges, and medians) appear in Tables 7 and 8 for the CST and WAB-R tasks, respectively. No significant systematic differences were obtained for all variables, indicating stability between test and retest results. The associations between test and retest were all significant, with strengths ranging from weak to moderate across variables.

Table 7. Descriptive statistics of the Core Lexicon variables for the Cinderella Story Retell task. Statistical testing used Wilcoxon signed-rank test for paired samples ('V' = test statistic; p = p value) comparing test and retest and Spearman's correlation assessing the strength of association between test and retest.

Variables	Test (n=48)		Retest (n=45)		Statistics		Interpretation
	Mean (SD)	Median [Min - Max]	Mean (SD)	Median [Min - Max]	V (*p* value)	Spearman' rho (*p* value)	
Core nouns	13.46 (3.29)	14 [4 – 18]	13.62 (3.08)	14 [4 – 18]	400.0 (*p*=.888)	0.360 (*p*=.015)*	No systematic difference, weak relationship between sessions.
Core verbs	11.46 (2.48)	12 [5 – 15]	11.38 (2.23)	12 [6 - 15]	372.0 (*p*=.983)	0.292 (*p*=.052)*	No systematic difference, weak relationship between sessions.
Core nouns per minute	4.93 (1.70)	4.78 [2.23 – 11.43]	5.08 (1.97)	4.71 [2.43 - 12.97]	535.0 (*p*=.641)	0.320 (*p*<.05)*	No systematic difference, moderate relationship between sessions.
Core verbs per minute	4.31 (1.88)	4.28 [1.84 – 14.29]	4.30 (1.72)	4.08 [2.13 - 11.14]	511.0 (*p*=.852)	0.592 (*p*<.001)*	No systematic difference, moderate relationship between sessions.

SD = Standard Deviation.
* Significant.

Table 8. Descriptive statistics of the Core Lexicon variables for the Picnic picture description of the WAB-R task. Statistical testing used Wilcoxon signed-rank test for paired samples ('V' = test statistic; *p* = *p* value) comparing test and retest and Spearman's correlation assessing the strength of association between test and retest.

Variables	Test (n=66)		Retest (n=66)		Statistics		Interpretation
	Mean (SD)	Median [Min - Max]	Mean (SD)	Median [Min - Max]	V (p value)	Spearman' rho (p value)	
Core nouns	14.33 (2.42)	15.0 [9 - 18]	13.83 (2.39)	14.0 [9 - 19]	414.0 ($p = .070$)	0.474 ($p < .001$)*	No systematic difference, moderate relationship between sessions.
Core verbs	4.08 (0.73)	4.0 [3 - 5]	3.97 (0.91)	4.0 [2 - 5]	497.50 ($p = .439$)	0.257 ($p < .05$)*	No systematic difference, weak relationship between sessions.
Core nouns per minute	11.99 (4.70)	11.32 [5.09 - 28.89]	11.36 (4.49)	10.63 [2.68 - 26.25]	887.0 ($p = .163$)	0.553 ($p < .001$)*	No systematic difference, moderate relationship between sessions.
Core verbs per minute	3.46 (1.49)	3.2 [1.19 - 7.05]	3.28 (1.40)	2.95 [0.77 - 7.50]	935.0 ($p = .276$)	0.530 ($p < .001$)*	No systematic difference, moderate relationship between sessions.

SD= Standard deviation.
* Significant.

A summary of long-term test-retest reliability and MDC90 results for all core lexicon variables for both tasks is presented in Table 9. MDC values were calculated for each variable to provide an indicator of clinical change. For example, the MDC90 for the core nouns list of the WAB-R was 3.75, meaning that a difference of 4 or more words would suggest a change attributable to other factors (e.g., language deterioration) rather than measurement error. ICC values and their corresponding confidence intervals ranged from poor to moderate for all variables, consistent with the significant but weak to moderate Spearman rho correlations. For the Cinderella Story Telling task, the best ICC results were obtained for the core lexicon verbs (ICC = 0.426, 95% CI [0.151, 0.639]) and the core lexicon verbs per minute (ICC = 0.519, 95% CI [0.266, 0.704]). For the Picnic picture description of the WAB-R task, the best ICC result was obtained for the core lexicon nouns (ICC = 0.521, 95% CI [0.323, 0.676]) and the core lexicon nouns per minute (ICC = 0.637, 95% CI [0.469, 0.639]).

Table 9. Summary of long-term test-retest results.
Koo and Li (2016) gives the following suggestion for interpreting intraclass correlation coefficient (ICC). including confidence intervals: below 0.50 = poor; between 0.50 and 0.75 = moderate; between 0.75 and 0.90 = good; and above 0.90 = excellent.

Measure	ICC	95% CI Low - High	Koo & Li (2016) ICC Quality (CI Quality)	Spearman' rho		Absolute Value Difference Between Test and Retest		MDC90
				r	*p* value	M (SD)	Range	
Cinderella Story Retell								
Core nouns	0.322	0.030 - 0.562	Poor (Poor - Moderate)	0.360	0.015	2.69 (2.61)	0 - 10	6.16
Core verbs	0.426	0.151 - 0.639	Poor (Poor - Moderate)	0.292	0.052	2.02 (1.48)	0 - 5	4.13
Core nouns per minute	0.160	-0.143 - 0.433	Poor	0.320	<.05	1.60 (1.75)	0 - 8.11	3.9
Core verbs per minute	0.519	0.266 - 0.704	Moderate (Poor - Moderate)	0.592	< 0.001	1.00 (1.47)	0 - 7.09	2.91
Picnic picture description of the WAB-R								
Core nouns	0.521	0.323 - 0.676	Moderate (Poor - Moderate)	0.474	< 0.001	1.62 (1.62)	0 - 6	3.75
Core verbs	0.253	0.013 - 0.465	Poor	0.257	<.05	0.80 (0.61)	0 - 3	1.66
Core nouns per minute	0.637	0.469 - 0.639	Moderate (Poor - Moderate)	0.553	< 0.001	2.02 (2.48)	0.10 - 11.11	6.46
Core verbs per minute	0.493	0.288 - 0.656	Poor (Poor - Moderate)	0.530	< 0.001	1.17 (0.88)	0.03 - 4.46	2.40

SD = Standard Deviation; CI = Confidence Interval; MDC90= Minimal Detectable Change at 90% confidence.

Discussion

The present study aimed to develop noun and verb core lexicon lists in Laurentian French for two discourse tasks—the Cinderella Story Telling (CST) and the Picnic picture description from the WAB-R—and to assess their psychometric properties. Following prior studies (Alyahya et al., 2021; Dalton & Richardson, 2015; Dalton et al., 2024), words were included in the core lexicon if they were produced by more than 50% of participants without brain injury (PWBI). This yielded four separate lists: 19 nouns and 5 verbs for the WAB-R, and 19 nouns and 16 verbs for the CST. We also assessed the rater and long-term test-retest reliability as well as construct validity for core lexicon variables. As expected, intra- and inter-rater reliability were good to excellent and construct validity analyses revealed significant positive correlations between core lexicon measures and micro- and macro-structural variables. There were no systematic differences and significant positive correlations between test and retest scores, suggesting general stability over time. However, ICC values remained poor to moderate, reflecting low consistency in individual rankings across timepoints—highlighting the paradox whereby stable group means may coexist with weak test–retest reliability (Hedge et al., 2018).

As expected, the inter-rater reliability results were excellent for all core noun lists and core verbs of the CST. Inter-rater reliability was good for the WAB-R core verbs, likely due to the limited score range resulting from the small number of target verbs (maximum score of five). The standard error measurements were also higher for the core verbs and nouns of the WAB-R task compared to the CST task, possibly reflecting the reliability paradox (Hedge et al., 2018), wherein restricted score variability can yield higher error estimates despite consistent scoring. Indeed, the SEM was higher in our participants than

in previous results with PWA (Kim & Wright, 2020). Higher SEM in PWBI compared to PWA has been previously reported for the microstructural variables of CIU and number of words per minute (Stark et al., 2023). The most likely explanation is that scoring tends to be more consistent and less variable in PWBI, particularly when the measure has a limited range of possible values (e.g., only five verbs in the WAB-R task), which can result in a higher standard error of measurement (SEM).

Construct validity results support using core lexicon nouns and verbs from both the CST and WAB-R tasks to assess lexical abilities. For the CST, core lexicon measures were significantly correlated with Main Concepts, mirroring findings by Dalton et al. (2015) in PWA. Our results also partially align with Kim and Wright (2020), who reported strong associations between core lexicon and key ideas in two storytelling tasks. However, while we observed significant correlations between core noun and verb scores and the total number of CIUs—a raw count reflecting informativeness—Kim and Wright found no significant associations when using the percentage of information units. This discrepancy likely reflects both differences in analytic approach (raw totals vs. percentages) and participant populations (PWBI vs. PWA), which can influence discourse variability and performance range.

For the WAB-R task, the macrostructural measure of Thematic Units correlated with core verbs and core nouns. In addition, the strongest correlations were obtained between core nouns per minute and Thematic Units per minute. We suggest that these results indicate a relationship between the ability to produce thematic content and essential lexical items efficiently. Indeed, it is unsurprising that using relevant and precise lexical units is associated with better results in conveying relevant global information about a stimulus.

A broader examination of our construct validity analyses revealed that core lexicon efficiency measures showed the strongest correlations with micro- and macro-structural measures for the WAB-R task. In contrast, for the CST task, total core lexicon scores demonstrated stronger and more consistent correlations with micro- and macro-structural variables than efficiency scores. We hypothesize that efficiency measures are more informative in shorter, highly constrained tasks such as the WAB-R picture description, where the brief sample length increases their sensitivity to lexical retrieval. By comparison, in the longer CST narratives, total core lexicon scores may better capture lexical performance across extended discourse.

To our knowledge, this study is the first to examine the test–retest reliability of core lexicon measures, including in languages other than English. Core lexicon variables have shown promise in differentiating groups at a single timepoint, including PWBI and individuals with cognitive decline (Fromm et al., 2024). Our findings extend this work by focusing specifically on their stability over time. In addition to ICCs, we reported Minimal Detectable Change (MDC90) values derived from the standard error of measurement. As highlighted by Boyle (2014) and Donoghue and Stokes (2009), MDC provides a clinically meaningful benchmark: it estimates the smallest change in a score that can be interpreted with confidence as a real change rather than measurement error. MDC90, based on a 90% confidence interval, is particularly recommended for evaluating change in individual performance and is valuable for monitoring outcomes in subclinical or longitudinal contexts. Across both discourse tasks, long-term test–retest analyses yielded poor to moderate intraclass correlation coefficients (ICCs), with no variable reaching the commonly accepted threshold for research applications (ICC > .70

Fitzpatrick et al., 1998), and none meeting the higher standard typically required for clinical use (ICC > .90). These results add to a mixed body of evidence on the test-retest reliability of discourse measures. For example, Stark et al. (2023) reported few systematic differences between test and retest when using short retest intervals, although their sample size (n = 24) was smaller than in the present study. Other studies employing longer intervals have also noted modest reliability estimates (Brisebois et al., 2023; Marcotte et al., 2024). Notably, variables such as mean length of utterance, noun/verb ratio (Brisebois et al., 2023; Stark et al., 2023), and Information Content Units (Marcotte et al., 2024) have similarly failed to demonstrate strong test–retest reliability. One likely explanation is that discourse production is inherently variable across timepoints, particularly in complex tasks like storytelling, where lexical choices and narrative structure may shift with each retelling (Fergadiotis & Wright, 2011). This may be especially true for PWBI, who might produce lexically rich but distinct samples upon repeated administration, thereby lowering score consistency despite intact discourse ability. In our study, reliability was highest for the WAB-R picture description—a more constrained task that elicited shorter, more uniform samples—where core nouns and core nouns per minute achieved moderate ICCs. While prior work suggests that longer samples improve reliability (Brookshire & Nicholas, 1994), our findings highlight that task structure and constraints may compensate for shorter output by reducing discourse variability. Future studies should examine how task type, discourse length, and lexical focus interact to influence the stability of discourse measures over time. Taken together, our results suggest that while core lexicon measures are valuable for assessing lexical performance at a single timepoint, they may not, in the absence of supporting evidence,

be optimal for capturing change over time. This aspect should be considered when applying core lexicon scoring to longitudinal designs or intervention studies.

Language and communication assessment standards in adults have greatly expanded in the last decade (Wallace et al., 2019). Even in discourse assessment, standards of reporting studies (Stark et al., 2022) and guidance for clinicians to assess discourse (Boyle, 2020) are now available. Nonetheless, discourse assessment in Laurentian French still faces numerous difficulties. To our knowledge, this is the first assessment of the core lexicon in French. Like Chan and Cheng (2024), the present results highlight the importance of studying discourse variables in different languages (García et al., 2023). Namely, our participants appeared to produce proportionally fewer unique lemmas than those reported in previous studies with English-speaking PWBI. While direct comparisons are limited by differences in sample size and task structure, the overall lexical diversity observed in both studies appears broadly comparable. This suggests that, despite linguistic and methodological differences, core lexicon measures may reflect similar discourse properties across languages. If we look at the most frequent words for our sample, we can find similarities but also discrepancies with previous lists. Indeed, six of our list's ten most frequent nouns are also on the early list produced by MacWhinney et al. (2010). Namely, Laurentian French equivalents of 'prince,' 'Cinderella,' 'fairy,' 'dress,' 'ball,' and 'shoe' were in our top ten nouns. Regarding verbs, eight of the ten most common verbs in English were also in the top ten in French. Unsurprisingly, the most frequent verb in French and English was 'to be' (i.e., 'être' in French), followed by 'to have' and 'to go' (respectively 'avoir' and 'aller' in French). Moreover, we must note that core verb scoring requires special attention since verb forms are more diverse than in

English. Indeed, French conjugations commonly imply root modification. For instance, the verb 'pouvoir' [can], could have the following forms: 'peut', 'peuvent', 'puisse', 'pu'. Despite this, core lexicon scoring demonstrated excellent intra- and inter-rater reliability. In summary, our investigation highlighted features of the French language, while also pointing to cross-linguistic regularities that reflect higher-order structures beyond language-specific differences. Also, to our knowledge, this study is the first to develop core lexicon lists for the WAB-R picture description task. Expanding core lexicon tools to include this widely used discourse task increases their applicability in clinical assessment and research.

Limitations

We acknowledge that typical test–retest reliability studies favor short intervals (~ 2 weeks) to reduce the influence of aging, learning, or other external factors. However, our use of a longer interval (~ 8 months) was a deliberate choice grounded in ecological validity. In clinical settings involving older adults or individuals at risk for neurodegenerative conditions, reassessments are often conducted months or even years apart. Previous research in healthy aging populations has demonstrated acceptable psychometric stability across similar timeframes, with intraclass correlations ranging from moderate to good over 8–13 months (e.g., Alioto et al., 2017). These longer intervals better reflect real-world follow-up scenarios such as cognitive monitoring and reduce inflation of reliability from practice effects. While our findings may not directly inform short-interval clinical retesting, they contribute necessary data on discourse in typical aging over clinically meaningful timelines. It is also important to consider the potential impact of sample size on our results. Compared to the groups assessed by

Dalton and Richardson (2015; n = 165 PWBI), Kim et al. (2019; n = 470 PWBI), Stark et al. (2023; n = 24), and Jiang et al. (2023; n = 88 PWBI), we included one group of 66 and the other of 48 PWBI. Hence, our sample sizes are more similar to the ones of Chen and Chang (2024; n = 43 PWBI) and Dalton and colleagues (2024; n = 45 and 50 PWBI).

Conclusion

This study is a first step towards a better understanding of core lexicon production and fundamental to supporting clinical implementation of core lexicon variables in Laurentian French PWBI. Because of its excellent intra- and inter-rater reliability and non-transcription-based analysis, core lexicon is very appealing to be transferred to clinical settings. However, our results suggest that test-retest reliability in target populations should be assessed before implementation. The current investigation supports future studies of core lexicon with participants with acquired communication difficulties, including aphasia and cognitive impairments affecting language.

Acknowledgments

We are very grateful to all the participants for their contribution to this study. This project was funded by the Heart and Stroke Foundation (grant-in-aid numbers G-16-00014039 and G-19-0026212) to K.M. and S.M.B. K.M. and S.M.B. hold a Career Award from the "Fonds de Recherche du Québec – Santé". A.B. holds a scholarship from the "Fonds de Recherche du Québec – Santé".

Data availability statement

The raw data presented in this article are not readily available because of the sensitivity of the video materials. The datasets analyzed during the current study are available from

the corresponding author upon reasonable request.

References

Alioto, A. G., Kramer, J. H., Borish, S., Neuhaus, J., Saloner, R., Wynn, M., & Foley, J. M. (2017). Long-term test-retest reliability of the California Verbal Learning Test – second edition. *The Clinical neuropsychologist*, *31*(8), 1449-1458. https://doi.org/10.1080/13854046.2017.1310300

Alyahya, R. S. W., Halai, A. D., Conroy, P., & Ralph, M. A. L. (2021). Content word production during discourse in aphasia : deficits in word quantity, not lexical–semantic complexity. *Journal of Cognitive Neuroscience*, *33*(12), 2494-2511. https://doi.org/10.1162/JOCN_A_01772

Antonsson, M., Lundholm Fors, K., Eckerström, M., & Kokkinakis, D. (2021). Using a discourse task to explore semantic ability in persons with cognitive impairment. *Frontiers in Aging Neuroscience*, *12*. https://www.frontiersin.org/articles/10.3389/fnagi.2020.607449

Armstrong, E. (2000). Aphasic discourse analysis : The story so far. *Aphasiology*, *14*(9), 875-892. https://doi.org/10.1080/02687030050127685

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. (2020). Choosing discourse outcome measures to assess clinical change. *Seminars in speech and language*, *41*(1), 1-9. https://doi.org/10.1055/S-0039-3401029

Brisebois, A., Brambati, S. M., Désilets-Barnabé, M., Boucher, J., García, A. O., Rochon, E., Leonard, C., Desautels, A., & Marcotte, K. (2020). The importance of thematic

informativeness in narrative discourse recovery in acute post-stroke aphasia. *Aphasiology*, *34*(4), 472-491. https://doi.org/10.1080/02687038.2019.1705661

Brisebois, A., Brambati, S. M., Jutras, C., Rochon, E., Leonard, C., Zumbansen, A., Anglade, C., & Marcotte, K. (2023). Adaptation and reliability of the Cinderella story retell task in Canadian French persons without brain injury. *American Journal of Speech-Language Pathology*, *32*, 2871-2888. https://doi.org/10.1044/2023_AJSLP-23-00101

Brookshire, R. H., & Nicholas, L. E. (1994). *Test-retest stability of measures of connected speech in aphasia* [Clinical Aphasiology Paper]. Clinical Aphasiology; Pro-Ed. http://aphasiology.pitt.edu/163/

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

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around : Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543-570. https://doi.org/10.1080/13854046.2012.680913

Chen, J., & Chang, H. (2024). Core Lexicon analysis of spoken discourse production by Mandarin Chinese speakers with Aphasia. *Aphasiology*. https://www.tandfonline.com/doi/abs/10.1080/02687038.2024.2325169

Colin, C., & Le Meur, C. (2016). *Adaptation du projet AphasiaBank à la langue française: Contribution pour une évaluation informatisée du discours oral de patients aphasiques* [Université Paul Sabatier, Toulouse III. Toulouse, France.]. http://thesesante.ups-tlse.fr/1747/

Covington, M. A. (2007). *MATTR user manual (CASPR Research Report 2007–05)*. University of Georgia Institute for Artificial Intelligence.

Dalton, S. G., AL Harbi, M., Berube, S., & Hubbard, H. I. (2024). Development of main concept and core lexicon checklists for the original and modern Cookie Theft stimuli. *Aphasiology*, *38*(12), 1975-1999. https://doi.org/10.1080/02687038.2024.2340794

Dalton, S. G., Kim, H., Richardson, J. D., & Wright, H. 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. D. (2015). Core-lexicon and main-concept production during picture-sequence description in adults without brain damage and adults with aphasia. *American Journal of Speech-Language Pathology*, *24*(4), S923-938. https://doi.org/10.1044/2015_AJSLP-14-0161

Deng, B.-M., Liang, L.-S., Zhao, J.-X., Zheng, H.-Q., & Hu, X.-Q. (2024). Correct Information Unit Analysis in Different Discourse Tasks Among Persons With Anomic Aphasia Based on Mandarin AphasiaBank. *American Journal of Speech-Language Pathology*, *33*(2), 800-813. https://doi.org/10.1044/2023_AJSLP-23-00217

Donoghue, D., Physiotherapy Research and Older People (PROP) group, & Stokes, E. K. (2009). How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. Journal of Rehabilitation Medicine, 41(5), 343-346. https://doi.org/10.2340/16501977-0337

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

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

Forbes-McKay, K. E., & Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences*, *26*(4), 243-254. https://doi.org/10.1007/s10072-005-0467-9

Frederiksen, C. H., & Stemmer, B. (1993). Conceptual processing of discourse by a right hemisphere brain-damaged patient. In H. Brownell & Y. Joanette (Éd.), *Narrative discourse in neurologically impaired and normal aging adults* (p. 239-278). Singular Publishing Group.

Fromm, D. A., Forbes, M., Holland, A., & MacWhinney, B. (2013). *PWAs and PBJs : Language for describing a simple procedure*. https://aphasiology.pitt.edu/2491/

Fromm, D., Dalton, S. G., Brick, A., Olaiya, G., Hill, S., Greenhouse, J., MacWhinney, B., & Nielson, K. (2024). The case of the Cookie Jar : differences in typical language use in dementia. *Journal of Alzheimer's Disease*, *100*(4), 1417-1434. https://doi.org/10.3233/JAD-230844

Fromm, D., Forbes, M., Holland, A., Dalton, S. G. H., 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

García, A. M., de Leon, J., Tee, B. L., Blasi, D. E., & Gorno-Tempini, M. L. (2023). Speech and language markers of neurodegeneration : A call for global equity. *Brain*, *146*(12), 4870-4879. https://doi.org/10.1093/brain/awad253

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox : Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166-1186. https://doi.org/10.3758/s13428-017-0935-1

Jiang, Y.-E., Liao, X.-Y., & Liu, N. (2023). Applying core lexicon analysis in patients with anomic aphasia : Based on Mandarin AphasiaBank. *International Journal of Language & Communication Disorders*, *58*(6), 1875-1886. https://doi.org/10.1111/1460-6984.12864

Kertesz, A. (2006). *Western Aphasia Battery- Revised*. Pearson.

Kim, B. S., & Lee, M. S. (2023). Discourse performance and related cognitive function in mild cognitive impairment and dementia : A preliminary study. *Applied Neuropsychology. Adult*, *30*(2), 194-203. https://doi.org/10.1080/23279095.2021.1922408

Kim, H., Berube, S., & Hillis, A. E. (2022). Core lexicon in aphasia: A longitudinal study. Aphasiology, 37(10), 1679–1691. https://doi.org/10.1080/02687038.2022.2121598

Kim, H., Kintz, S., & Wright, H. H. (2020). Development of a measure of function word use in narrative discourse : Core lexicon analysis in aphasia. *International Journal of Language and Communication Disorders*. https://doi.org/10.1111/1460-6984.12567

Kim, H., Kintz, S., Zelnosky, K., & Wright, H. H. (2019). Measuring word retrieval in narrative discourse : Core lexicon in aphasia. *International Journal of Language and Communication Disorders*, *54*(1), 62-78. https://doi.org/10.1111/1460-6984.12432

Kim, H., Sung, J. E., & Jeong, J. H. (2022). Non-transcription analysis of connected speech in mild cognitive impairment using an information unit scoring system. *Journal of Neurolinguistics*, *61*, 101035. https://doi.org/10.1016/j.jneuroling.2021.101035

Kim, H., & Wright, H. H. (2020). 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

Kintz, S., Kim, H., & Wright, H. H. (2024). A preliminary investigation on core lexicon analysis in dementia of the Alzheimer's type. *International Journal of Language & Communication Disorders*, *59*(4), 1336-1350. https://doi.org/10.1111/1460-6984.12999

Koo, T. K., & Li, M. Y. (2016). A guideline of 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

MacWhinney, B. (2000). *The CHILDES Project : Tolls for Analzying Talk: Vol. 3rd Editio*. Lawrence Erlbaum Associates.

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

Marcotte, K., Lachance, A., Brisebois, A., Mazzocca, P., Désilets-Barnabé, M., Desjardins, N., & Brambati, S. M. (2022). Validation of videoconference administration of picture description from the Western Aphasia Battery–Revised in neurotypical Canadian French speakers. *American Journal of Speech-Language Pathology*, *31*(6), 2825-2834. https://doi.org/10.1044/2022_AJSLP-22-00084

Marcotte, K., Roy, A., Brisebois, A., Jutras, C., Leonard, C., Rochon, E., & Brambati, S. M. (2024). Reliability of the picture description task of the Western Aphasia Battery – revised in Laurentian French persons without brain injury. *The Clinical Neuropsychologist*, *0*(0), 1-29. https://doi.org/10.1080/13854046.2024.2340777

Matthews, P. H. (2014). Lemma. In *The Concise Oxford Dictionary of Linguistics*. Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780199675128.001.0001/acref-9780199675128-e-1845

Mueller, K. D., Hermann, B., Mecollari, J., & Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease : A review of picture description tasks. *Journal of clinical and experimental neuropsychology*, *40*(9), 917-939. https://doi.org/10.1080/13803395.2018.1446513

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research*, *36*(2), 338-350. https://doi.org/10.1044/jshr.3602.338

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2017). Reviewing the quality of discourse information measures in aphasia. *International Journal of Language & Communication Disorders*, *52*(6), 689-732. https://doi.org/10.1111/1460-6984.12318

Schnur, T. T., & Wang, S. (2024). Differences in connected speech outcomes across elicitation methods. *Aphasiology*, *38*(5), 816-837. https://doi.org/10.1080/02687038.2023.2239509

Slegers, A., Filiou, R.-P., Montembeault, M., & Brambati, S. M. (2018). Connected speech features from picture description in Alzheimer's disease : a systematic review. *Journal of Alzheimer's Disease*, *65*(2), 519-542. https://doi.org/10.3233/JAD-170881

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. https://archive.mpi.nl/tla/elan

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

Stark, B. C., Bryant, L., Themistocleous, C., den Ouden, D.-B., & Roberts, A. C. (2022). Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology*, *0*(0), 1-24. https://doi.org/10.1080/02687038.2022.2039372

Taler, V., Davidson, P. S. R., Sheppard, C., & Gardiner, J. (2021). A discourse-theoretic approach to story recall in aging and mild cognitive impairment. *Aging, Neuropsychology, and Cognition*, *28*(5), 762-780. https://doi.org/10.1080/13825585.2020.1821865

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Breitenstein, C., Hilari, K., Babbitt, E., Bose, A., Brady, M., Cherney, L. R., Copland, D., Cruice, M., Enderby, P., Hersh, D., Howe, T., Kelly, H., Kiran, S., Laska, A.-C., Marshall, J., … Webster, J. (2019). A core outcome set for aphasia treatment research : The ROMA consensus statement. *International Journal of Stroke: Official Journal of the International Stroke Society*, *14*(2), 180-185. https://doi.org/10.1177/1747493018806200

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., Miller, B. L., & Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain*, *133*(7), 2069-2088. https://doi.org/10.1093/brain/awq129

Supplemental Material S1. Best Practice Guidelines for Reporting Spoken Discourse in Aphasia and Neurogenic Communication Disorders.

Supplemental Material S2. Core lexicon scoring template.