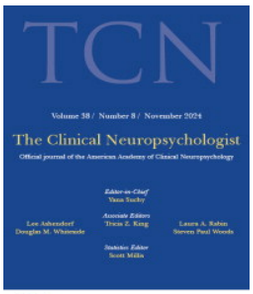

The Clinical Neuropsychologist

ISSN: 1385-4046 (Print) 1744-4144 (Online) Journal homepage: www.tandfonline.com/journals/ntcn20

Reliability of the picture description task of the Western Aphasia Battery – revised in Laurentian French persons without brain injury

Karine Marcotte, Alexandra Roy, Amélie Brisebois, Claudie Jutras, Carol Leonard, Elizabeth Rochon & Simona Maria Brambati

To cite this article: Karine Marcotte, Alexandra Roy, Amélie Brisebois, Claudie Jutras, Carol Leonard, Elizabeth Rochon & Simona Maria Brambati (2024) Reliability of the picture description task of the Western Aphasia Battery – revised in Laurentian French persons without brain injury, The Clinical Neuropsychologist, 38:8, 1980-2008, DOI: 10.1080/13854046.2024.2340777

To link to this article: https://doi.org/10.1080/13854046.2024.2340777

THE CLINICAL NEUROPSYCHOLOGIST
2024, VOL. 38, NO. 8, 1980–2008
https://doi.org/10.1080/13854046.2024.2340777

Routledge
Taylor & Francis Group

Reliability of the picture description task of the Western Aphasia Battery – revised in Laurentian French persons without brain injury

Karine Marcotte[a,b], Alexandra Roy[a], Amélie Brisebois[a,b], Claudie Jutras[b,c], Carol Leonard[d,e,f], Elizabeth Rochon[e,f,g,h] and Simona Maria Brambati[b,c,i]

[a]École d'orthophonie et d'audiologie, Faculté de médecine, Université de Montréal, Montréal, Québec, Canada; [b]Centre de recherche du Centre intégré universitaire de santé et de services sociaux du Nord-de-l'Île-de-Montréal, Montréal, Québec, Canada; [c]Département de psychologie, Faculté des arts et des sciences, Université de Montréal, Montréal, Québec, Canada; [d]School of Rehabilitation Sciences, University of Ottawa, Ontario, Ottawa, Canada; [e]Department of Speech-Language Pathology, Temerty Faculty of Medicine, University of Toronto, Toronto, Canada; [f]Heart and Stroke Foundation, Canadian Partnership for Stroke Recovery, Ontario, Canada; [g]KITE Research Institute, Toronto Rehabilitation Institute, Toronto, Canada; [h]Rehabilitation Sciences Institute, University of Toronto, Toronto, Canada; [i]Centre de recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, Québec, Canada

ABSTRACT

Objective: Limited normative data (including psychometric properties) are currently available on discourse tasks in non-dominant languages such as Laurentian (Quebec) French. The lack of linguistic and cultural adaptation has been identified as a barrier to discourse assessment. The main aim of this study is to document inter-rater and test-retest reliability properties of the picnic scene of the Western Aphasia Battery – Revised (WAB-R), including the cultural adaptation of an information content unit (ICU) list, and provide a normative reference for persons without brain injury (PWBI). **Method:** To do so, we also aimed to adapt an ICU checklist culturally and linguistically for Laurentian French speakers. Discourse samples were collected from 66 PWBI using the picture description task of the WAB-R. The ICU list was first adapted into Laurentian French. Then, ICUs and thematic units (TUs) were extracted manually, and microstructural variables were extracted using CLAN. Inter-rater reliability and test-retest reliability were determined. **Results:** Excellent inter-rater reliability was obtained for ICUs and TUs, as well as for all microstructural variables, except for mean length of utterance, which was found to be good. Conversely, test-retest reliability ranged from poor to moderate for all variables. **Conclusion:** The present study provides a validated ICU checklist for clinicians and researchers working with Laurentian French speakers when assessing discourse with the picnic scene of the WAB-R. It also addresses the gap in available psychometric data regarding inter-rater and test-retest reliability in PWBI.

ARTICLE HISTORY

Received 17 December 2023
Accepted 4 April 2024
Published online 12 April 2024

KEYWORDS

Discourse analysis; reliability; picture description; persons without brain injury

CONTACT Karine Marcotte karine.marcotte@umontreal.ca 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 H4J 1C5, Canada.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/13854046.2024.2340777.

Introduction

Expressive discourse is fundamental for daily communication. Every day, we are called upon to produce discourse to tell the story of our day, to share an opinion on different subjects or simply to converse with others. These discourse skills come naturally and effortlessly to most of us. Although discourse production seems relatively easy, it involves a complex interplay of multiple language, cognitive and socio-demographic variables. Compared to single word production tasks, spoken discourse assessment thus offers a more ecological assessment of language impairments (Bryant et al., 2016; Stark et al., 2021). According to broad scientific consensus, discourse is defined as larger than an utterance or a sentence (Kong, 2016). In fact, discourse is the most elaborate manifestation of human expressive language (Ska et al., 2004). Discourse effectively allows for the examination of multiple language characteristics in much more natural contexts than other language tasks that have been more widely studied to date, such as picture naming (Prins & Bastiaanse, 2004), which requires only the production of single words. Therefore, a growing body of research has focused on spoken discourse assessment and analysis in post-stroke aphasia (Stark et al., 2021), and more recently in neurocognitive disorders such as Alzheimer's disease (Filiou et al., 2020; Mueller et al., 2018; Slegers et al., 2018). Discourse analysis is especially useful because it allows the simultaneous assessment of several functions, including the different language levels and other cognitive functions such as executive functions, in a more ecological way than tests targeting each function separately (Filiou et al., 2020).

Importance of discourse assessment in clinical settings

In a recent survey, 86% of speech-language pathologists reported that they performed discourse assessment in people with acquired communication impairment (Bryant et al., 2017). Single-picture description is most widely used in both persons without brain injury (PWBI) and in clinical populations (Bryant et al., 2016) and for both clinical and research purposes because it captures a wide range of information about language content, structure, and pragmatic skills in a relatively quick and easy task. Moreover, picture description tasks provide good ecological validity compared to single word elicitation tasks (Ahmed et al., 2013; Cooper, 1990; Doyle et al., 1995; Giles et al., 1996; Slegers et al., 2018). Picture description reduces cognitive demands on attention and executive functions (Giles et al., 1996; Slegers et al., 2018) as well as episodic memory because the story is visually presented to the participant (Duong et al., 2003). These tasks also offer a structured context with specific and restrained content, which allows clinicians and researchers to compare between individuals at different points in time (Boucher et al., 2022; Bryant et al., 2016; Chenery & Murdoch, 1994; Mackenzie et al., 2007).

Changes in discourse production can be observed in a variety of acquired neurogenic disorders such as traumatic brain injury, stroke, mild cognitive impairment (MCI), Alzheimer disease (AD), and primary progressive aphasia (PPA). Some changes can simply reflect the normal aging trajectory (Boschi et al., 2017; Capilouto et al., 2016; Filiou et al., 2020; Hillis, 2007; Le Dorze & Bédard, 1998; Mueller et al., 2018). Studies

examining discourse of individuals presenting language impairments associated with cognitive decline (e.g. MCI, AD) also mostly use a single image for picture description (Filiou et al., 2020). In their review of picture description tasks, Mueller et al. (2018) reported that semantic content, which can be examined by using thematic units (TUs) or relevant information content units (ICUs), have proven to be the most effective measures in capturing language deterioration in MCI and AD. Their review points out that robust observations about language impairment have been made in the latest stage of AD but there are still many aspects to explore to detect subtle preclinical changes in discourse in early cognitive decline.

More recently, a group of researchers has proposed the definition of *subjective cognitive decline* (SCD). The SCD criteria include two main features: a) a self-reported persistent cognitive decline without evidence of an acute event, and b) normal performance using standardized objective tests (Jessen et al., 2020). Studies suggest that SCD could foreshadow future deterioration of cognitive functions (Jessen et al., 2014; Mitchell et al., 2014; Slot et al., 2019). Several cognitive domains can be affected in SCD, including language. Verfaillie et al. (2019) reported that the use of specific words produced in discourse was associated with high levels of amyloid burden in individuals with SCD, whereas no conventional neuropsychological language tests nor other discourse measure found such association. Profiling discourse feature trajectories would be useful to capture subtle changes across cognitive decline and allow early recognition of these individuals. Considering that SCD is usually not detected by standard cognitive testing, its identification requires measures highly sensitive and with robust psychometrical features (Jessen et al., 2014).

Methodological challenges in discourse analysis

There is a consensus that it is almost mandatory to diversify sampling methods and carefully select analysis procedures to obtain a representative picture of discourse skills (Bryant et al., 2016), but no consensus on which measures and tasks should be used has yet arisen (Dietz & Boyle, 2018). Regarding the task itself, the selected task, or set of tasks, can create disparities among two persons from different cultures. For instance, some tasks, such as the Cookie Theft (Goodglass et al., 2001), were developed decades ago and depict a scene from the past century including cultural, linguistic and socioeconomic bias (Steinberg et al., 2022). Other tasks tend to be more inclusive of multicultural individuals, but few have investigated the multicultural impact of the stimuli on performance until recently, with the precarious painter scene (Stockbridge et al., 2024). Moreover, language performance, in terms of content and length of productions, can vary depending on the task selected to elicit spoken discourse (Boucher et al., 2022; Bryant et al., 2016). Several picture description tasks are available, but the visual elements of the pictorial stimuli (e.g. number of elements, spatial location of the elements, relationships between the elements) are highly variable, which may in turn affect production. The choice of task is thus crucial because it must be socially and culturally adapted to provide a representative sample of discourse production.

The choice of the measures extracted can also affect results obtained in the different studies. The large methodological differences across studies with regards to

the discourse measures constitute major challenges for researchers when comparing results across studies (Dietz & Boyle, 2018), as well as for clinicians when selecting outcome measure(s) (Azios et al., 2022). In a review of 165 studies focusing on linguistic discourse analysis of people with aphasia, Bryant et al. (2016) reported a total of 536 different linguistic measures for language analysis. To date, most studies that have conducted discourse analysis have focused on the macrostructural and microstructural variables of discourse, which are composed of the two first stages (i.e. conceptual preparation and linguistic formulation) of Frederiksen's model of discourse (Frederiksen & Stemmer, 1993). Macrostructural measures refer to a higher-level conceptual structure of discourse (Dijk, 2019), such as informativeness, coherence, and cohesion. Among the most studied macrostructural variables, informativeness assesses the ability of an individual to convey relevant information about a given stimulus (Armstrong, 2000). A variety of measures have been used to examine informativeness, such as content units (also called by others information content unit (ICU); Yorkston & Beukelman, 1980), main concept analysis (MCA; Nicholas & Brookshire, 1995), and more recently, thematic units (TUs; Marini et al., 2011). An ICU quantifies key elements in a pictorial stimulus which can be divided into different categories (e.g. objects, people, places, and actions). The TU checklist, on the other hand, is based on a finite set of semantic or more general themes, which may arguably increase its reliability (Brookshire & Nicholas, 1994). One of the main advantages of ICUs and TUs is that they are easy and quick to score, which increases their applicability in clinical settings. However, the reliability of these measures requires further investigation with larger sample sizes.

On the other hand, microstructural measures refer to local or within-sentence features involving phonological, lexical, semantic and grammatical processing. Mean length of utterance (MLU), duration, number of words per minute (WPM) and moving average token-type ratio (MATTR) have been shown to be the most sensitive to language impairment. For instance, Brisebois et al. (2023) found that multilevel analysis of discourse changes revealed a different evolution of variables at each discourse level in people with acquired communication impairments (e.g. Brisebois et al., 2023; Marini et al., 2011), which supports the importance of developing reliable discourse measures at both the macrostructural and microstructural levels.

Moreover, a recent international survey identified that the scarcity of discourse protocols and normative data, including psychometric properties, is a barrier to discourse assessment (Stark et al., 2021). In addition, test-retest reliability and inter-rater reliability have been reported for only a minority of measures (Pritchard et al., 2017). Test-retest reliability evaluates the consistency and the stability of a measure where participant behavior is tested with the same method, after a certain time interval (Schiavetti et al., 2011). Various intraindividual factors (e.g. tiredness, level of attention, etc.) have an impact on discourse production and impact day-to-day performance (Spencer et al., 2020). Documentation about the reliability of a measure throughout a certain period of time would guide individual clinical decision-making in differentiating between natural variation and a therapeutic effect (Brookshire & Nicholas, 1984). Also, exploring test-retest reliability of discourse measures could help make more informed assumptions about discourse trajectories in

normal aging and in people presenting cognitive decline, such as in SCD (Mueller et al., 2018). Therefore, natural intraindividual fluctuations found in discourse of the elderly are important to document. To our knowledge, most studies that have reported test-retest reliability of discourse metrics have done so using short intervals (i.e. between one and two weeks) in order to obtain reliable measures in the context of potential learning between two assessments (Bartels et al., 2010). However, a longer period between testing sessions may better reflect changes associated with typical aging (Mueller et al., 2018). Among the few studies focused on test-retest reliability, Boyle (2015) reported poor test-retest reliability when multiple discourses tasks were analyzed separately and showed an increase in stability over time of selected measures when various narrative tasks were combined. Similarly, Brookshire and Nicholas (1994) suggested that test-retest reliability can be improved by using multiple stimuli, or by increasing the sample size. These results suggest that clinicians and researchers should not draw conclusions based on a single picture description task. However, Stark et al. (2023) reported that test-retest reliability varied among the different tasks, which argues in favor of not combining different types of discourse.

As mentioned above, inter-rater reliability, is another important psychometric property to report. It evaluates the consistency of a score on the same samples by different raters. The recent review of Pritchard et al. (2017) indicated that inter-rater reliability was reported for approximately a third of discourse measures used. More importantly, the studies reviewed did not employ appropriate statistical methods to test reliability.

These results combined support the importance of studying the quality of measurements in terms of psychometric properties (Bryant et al., 2016; Dietz & Boyle, 2018; Linnik et al., 2016; Mueller et al., 2018; Pritchard et al., 2017, 2018; Simmons-Mackie & Lynch, 2013; Stark et al., 2021; Stark et al., 2022) for each group of participants, for each elicited task and for longer intervals considering that test-retest data currently available are not adapted for longitudinal studies (Mueller et al., 2018). The investigation of reliability of various discourse measures will help identify discourse measures with the best psychometrics properties for both research and clinical purposes.

The lack of linguistic and culturally adapted methods was an additional barrier in non-dominant languages, according to the previously mentioned international survey (Stark et al., 2021). Language(s) spoken by an individual can also impact language production profiles (Filiou et al., 2020; Mehler, 1994). Currently, we observe an over-representation of English-speakers in data available on language. This lack of language diversity in the languages investigated constitutes a barrier toward the development of globally equitable measures of connected speech and early identification of neurocognitive disorders such as AD (García et al., 2023). Research samples collected to date are not always representative of linguistic and cultural differences that constitute language diversity on a larger scale. Compared to the English-speaking population, the scope of assessments is more limited for French speakers, especially from the province of Quebec. French is not only a non-dominant language in Canada, but across North America. Many linguistic challenges are present considering that Quebec abounds in a unique linguistic richness because of its regional variants of the French language (i.e. dialects) and the presence of multilingualism. Over the

last few years, our team has focused on the standardization of discourse assessment in Laurentian (also known as Canadian or Quebec) French (Boucher et al., 2022; Brisebois et al., 2023; Marcotte et al., 2022). The present study is an extension of this work.

Aims of the study

The current study is an extension of our previous study (Boucher et al., 2022) that aimed to provide reference data for picture description of the picnic scene of the WAB-R (Kertesz, 2006) for adults over 50 years old. The main aim of the present study is to investigate the reliability of discourse measures at the micro- and macro-structural levels of discourse for the WAB-R picture description task, especially test-retest reliability, which was not tested in our previous study (Boucher et al., 2022). To do so, we also needed to develop a culturally and linguistically adapted list of ICUs. As recently reported by others (Stark et al., 2023) and our team (Brisebois et al., 2023), we expect good inter-rater reliability (IRR), but lower test-retest reliability in PWBI. Secondly, this study will provide reference data for the picnic scene of the WAB-R for Laurentian French PWBI.

Methods and materials

All necessary and recommended standards for reporting spoken discourse are reported in the manuscript. For more details, the best practice guidelines checklist from Stark et al. (2022) is provided in Supplementary Material 1.

Participants

The sample consisted of a subset of individuals from a previously published study (Marcotte et al., 2022). Briefly, 66 PWBI were recruited 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)* that aimed to investigate longitudinal post-stroke aphasia recovery. Eighteen participants were recruited for a project which sought to investigate longitudinal changes in post-stroke aphasia (CIUSSS-NIM; # MP-32-2018-1478). Another 48 participants were recruited during the COVID-19 pandemic for a project which sought to investigate longitudinal spoken discourse changes following a stroke (CIUSSS-NIM # 2020-1900). Written informed consent was obtained from all participants. The inclusion criteria for this study were: 1) to be at least 50 years of age; 2) have Laurentian (Quebec) French as their primary language of use at the time of the study. The exclusion criteria for this study were: 1) presenting a severe mental illness; 2) presenting an acquired or developmental language impairment; 3) suffering from a neurological impairment, including a neurocognitive impairment; 4) having suffered from a traumatic brain injury; 5) self-reporting cognitive decline or complaints; 6) uncorrected visual or auditory deficits. Exclusion criteria were assessed using a self-reported questionnaire completed by each participant prior to the study. Participant characteristics appear in Table 1. All participants were Caucasian.

Table 1. Participants' characteristics.

Variable	
Age	
Mean (SD)	64.53 (7.15)
Median [Min–Max]	64 [52–82]
Gender	
Female	37 (56.06%)
Male	29 (43.94%)
Handedness	
Right	60 (93.94%)
Left	4 (6.06%)
Education	
Mean (SD)	16.11 (2.86)
Median [Min–Max]	16 [11–25]
Time between sessions (days)	
Mean (SD)	253.36 (67.45)
Median [Min–Max]	252 [162–406]
Linguistic profile	
Monolingual (French only)	25 (37.88%)
Bilingual (French and English)	35 (53.03%)
Multilingual (French, English and other language(s))	6 (9.09%)

n = 66; SD = Standard Deviation; Min = Minimum; Max = Maximum.

Adaptation of the information content unit (ICU) list in Laurentian French

An ICU list of the picnic scene of the WAB-R (Kertesz, 2006) was originally developed for American English speakers (Jensen et al., 2006), and cultural adaptation requires that the target population shares a similar cultural background with the initial sample. Cultural and linguistically valid adaptations usually involve modifications, i.e. developing an entirely new task (Kong, 2009) or refining the scoring protocol (Brisebois et al., 2023; Criel et al., 2021; Yazu et al., 2022). Considering that Laurentian French speakers share a similar cultural background with American English speakers regarding the picnic scene, an adaptation was made by refining the scoring protocol. Thus, the ICU checklist was translated and adapted from the original list of Jensen et al. (2006). First, we used the online free version of DeepL Translator (*DeepL Traduction—DeepL Translate*, 2022) to translate the first draft of the 36 ICUs in French. Second, a research assistant (C.J.), who is a native Laurentian French speaker with advanced knowledge of written English, reviewed the first draft to ensure that each element was as semantically similar as possible to the original version as possible. Third, final adjustments were made *via* discussion between the research assistant, the principal investigator (K.M.) and a Ph.D. student (A.B.). Based on these discussions, two of the ICUs were combined ("On the beach" and "In the sand") because they are used interchangeably in French. Then, we compared the list with the one used in other studies (Boucher et al., 2022; Gallée et al., 2021). As a result, we added one ICU ("*run/is chasing*") to the action category considering the frequent production of this element in these studies. The final integrated translation of the ICU list is reported in the Results section.

Data collection

All participants completed a variety of tasks evaluating different language components, including the picnic scene of the WAB-R (Kertesz, 2006), which was the sole

discourse task. Tasks were completed twice, with the mean number of days between sessions equaling 253.36 ± 67.45 days and a range of 162–406 days. Audio recordings were collected for 18 participants who completed the task in person using a Sony IC recorder icd-px312 for 27 participants and a Sony HDR-PJ540 camera (9.2 megapixels). Discourse samples from the picnic scene were collected by video recording using the Zoom platform (https://zoom.us) for 48 participants. For the in-person group, the picnic scene stimulus was placed on the desk in front of the participant. For the videoconference group, further details regarding the procedure can be found in the Supplemental Material S1 of Marcotte et al. (2022). No significant difference has been found between in-person and videoconference administration of this task (Marcotte et al., 2022), which supports combining both groups in the present study.

Briefly, the task was administered by either trained research assistants or trained certified speech-language pathologists. Participants were asked to describe what they saw in the picture, using complete sentences (« *Décrivez en détail tout ce qui se passe sur cette image en utilisant des phrases complètes.* »). No time limit was given. If participants remained silent for more than 10 s, the examiner asked them once if they had anything else to add before ending the recording.

Transcription

The procedure for transcription was previously reported in Brisebois et al. (2020). Participants' discourse was transcribed verbatim. The Code for the Human Analysis of Transcripts (CHAT) manual (MacWhinney, 2000) was used for the phonemic transcription, utterance segmentation, transcription and scoring, with additional guidance for French speakers (Colin & Le Meur, 2016) and from the phonological, syntactic and semantic criteria proposed by Marini et al. (2011). Video recordings were imported and transcribed in the EUDICO Language Annotator (ELAN; Sloetjes & Wittenburg, 2008) by a trained research assistant or by an experienced speech-language pathologist. Once the transcription was completed, the morphological and grammatical information coding was conducted using the CLAN program called *mor* (MacWhinney, 2000), which tags morphemes and words under each utterance in the transcripts. Microstructural measures (described in Table 2) were extracted automatically from each sample at each time point using the EVAL program of CLAN software (MacWhinney, 2000 version of January 5, 2021, updated September 30, 2021).

Dependent variables

Discourse measures were selected based on previously reported research into discourse impairment associated with cognitive decline in people with neurocognitive disorders (Filiou et al., 2020; Slegers et al., 2018). Both macrostructural and microstructural variables are described in Table 2. All microstructural variables were extracted for each sample using the program EVAL of CLAN. Specific CLAN commands for each variable are provided in Table S1 of Supplementary Material 2.

Table 2. Definition of the discourse variables.

Measure	Definition	Language dimension
Macrostructural variables		
ICU_{total}	Total number of ICUs produced	General informativeness
ICUs per minute (ICUs/min)	Total number of ICUs divided by the duration (converted from seconds to minute)	General informativeness
ICUs per utterance	Total number of CIUs divided by the number of utterances	General informativeness
$ICU_{subjects}$	Total number of ICUs from the subject category produced	General informativeness
ICU_{places}	Total number of ICUs from the places category produced	General informativeness
$ICU_{entities}$	Total number of ICUs from the entities category produced	General informativeness
$ICU_{actions}$	Total number of ICUs from the action category produced	General informativeness
TU_{total}	Total number of TUs produced	Thematic informativeness
TUs per minute (TUs/min)	Total number of TUs divided by the duration (converted from seconds to minute)	Thematic informativeness
TUs per utterance (TUs/utterance)	Total number of TUs divided by the number of utterances	Thematic informativeness
Microstructural variables		
Duration	Duration of the sample in seconds	Corpus size
Tokens	Total number of words produced	Corpus size
Mean length of utterance (MLU)	Average number of words per utterance	Productivity
Propositional density	Number of verbs, adjectives, adverbs, prepositions and conjunctions divided by the total number of words	Content richness
Words per minute (WPM)	Total number of tokens divided by the duration (converted from seconds to minute)	Fluency
Verbs per utterance	Average number of verbs (verbs, copulas, auxiliaries followed by past or present participles) per utterance.	Syntactic complexity
Open/closed class ratio	Ratio of open class words (all nouns, verbs, copulas, adjectives and adverbs) divided by closed class words (all other words)	Syntactic complexity
Noun/verb ratio	Ratio of nouns to verbs, excluding auxiliaries and modals	Syntactic complexity
Moving Average Token-Type Ratio (MATTR)	Average of estimated Token-Type Ratios for successive nonoverlapping successive windows of flixed length	Lexical diversity
% Correct information units (CIUs)	Total number of words relevant to the stimulus and informative (CIUs) divided by the total number of words	Lexical informativeness
CIUs per minute (CIUs/min)	Total number of CIUs divided by the duration (converted from seconds to minute)	Lexical informativeness

Note. Data derived from the CLAN software (MacWhinney et al., 2011).

Data analysis

Analysis of ICU frequency

Previous test adaptation in Laurentian French has demonstrated cultural differences in performance on specific task items (e.g. Brisebois et al., 2023; Callahan et al., 2010). Hence, the frequency of each ICU was computed at test and retest. Only the ICUs which were produced by a minimum of 20% of the sample, as in Jensen et al. (2006), were kept in the final adaptation of the ICU checklist.

Inter-rater reliability

To determine inter-rater reliability in transcription, 15 transcripts (representing 11% of the transcripts) were randomly selected for a second transcription. Inter-rater reliability was computed for 3 variables: tokens, total number of utterances and CIUs. The total number of tokens represents the accuracy of the transcription. The number of utterances is critical in CHAT format since it relies uniquely on the transcriber's competence to distinguish utterance boundaries. Reliability on this measure suggests consistency in utterance segmentation throughout the samples. As for CIUs, they have

been more extensively studied in English (Fergadiotis et al., 2019), but have only been studied with the Cinderella story retell task in Laurentian French (Brisebois et al., 2023). To determine inter-rater reliability for scoring, 30 transcripts per rater (representing 22% of the transcripts) were selected randomly for two raters as before. Both raters scored the ICU and TU lists. A greater proportion of transcripts were selected since these measures have been less extensively studied.

Two-way mixed intraclass correlation coefficients (ICCs) with absolute agreement with a 95% confidence interval (CI) were calculated on both transcription (i.e. number of tokens, utterances and CIUs) and scoring variables (i.e. TUs and ICUs). Use of ICC for this purpose is optimal since this analysis takes into account absolute agreement and intra-group variability (Koo & Li, 2016). The interpretation of ICC values is based on guidelines reported in Koo and Li (2016): poor ($r < 0.50$), moderate ($0.50 < r < 0.75$), good ($0.75 < r < 0.90$) and excellent ($r > 0.90$).

Test-retest reliability

Data distribution was analysed using Kolmogorov-Smirnov test for all dependent variables, for each session. Consistent with similar studies (Stark et al., 2023), more than 70% of the data were not normally distributed. Consequently, we chose non-parametrical statistical analyses for all variables to maintain consistency. The Wilcoxon signed rank-test was used to determine if there was a difference between the two sessions for each discourse variable. Twenty-two comparisons were made; using the Bonferroni correction, alpha was set at .002. Spearman Rho correlations were used to assess the association between test and retest, with significance set at $p < 0.05$. Two-way mixed effects intra-class correlation (ICC) based on single measurement and absolute agreement with a 95% confidence interval (CI) were computed to evaluate test-retest reliability. As for inter-rater reliability, the interpretation of ICC values is based on guidelines reported in Koo and Li (2016).

Regarding agreement, visual inspection of the data was completed by examining the limits of agreement between testing points with Bland-Altman plots (Altman & Bland, 1983). Bland-Altman plots are scatterplots with the Y axis representing the difference between the results obtained at test and retest and the X axis representing the mean of the test and retest results. Limits of agreement are represented with horizontal dashed lines at ±1.96 standard deviations of the mean of differences. If 95% of the data falls between these limits, the agreement between test and retest is considered good (Bland & Altman, 1999). These plots were created for the variables that obtained the best test-retest ICC.

As in Stark et al. (2023), minimal Detectable Change (MDC) was also computed across all dependent variables using the standard error of measurement (SEM). The SEM formula includes standard derivation of tests (SDx) and correlation coefficient (rxy): $SEM = SD\sqrt{1\text{-}r}$. MDC is a well-known measure commonly used to investigate the variability in a score that reflects "real" change, greater than the measurement error. We also established a 90% confidence of prediction for MDC to estimate the possible change related to therapeutic gains (Donoghue & Stokes, 2009) or pathological change in cases of PWBI. The formula to calculate MDC90 is $MDC90 = SEM * 1.65 * \sqrt{(2)}$.

Analysis software

All statistical analyses were performed using SPSS® v26.0. Bland-Altman plots were computed using RStudio 2022.07.2.

Results

Development of the adapted ICU list

The frequency of each ICU was computed at test and retest and appears in Table 3. All ICUs reached the 20% frequency threshold used by Jensen et al. (2006) at both timepoints, except for one action (i.e. "*le drapeau* vole" [flag flies]) which reached 24% at test but 17% at retest. The action was kept in the final list because its mean frequency score was slightly above the 20% cut-off. The final list of ICUs adapted in Laurentian French with the detailed scoring guide is available in an Excel sheet "*Modèle à copier*" (i.e. template) in Supplementary Material 3.

Inter-rater reliability

Scoring reliability was excellent for both ICUs ($ICC_{[2,1]}$ = 0.973, 95% CI [0.944, 0.987]) and TUs ($ICC_{[2,1]}$ = 0.958, 95% CI [0.914, 0.991]). Transcription reliability was excellent for tokens ($ICC_{[2,1]}$ = 0.959, 95% CI [0.881, 0.986]) and the total number of CIU ($ICC_{[2,1]}$ = 0.988, 95% CI [0.932, 0.997]), and good for utterances ($ICC_{[2,1]}$ = 0.831, 95% CI [0.559, 0.941]). Detailed results are reported in Table S2 of Supplementary Material 2.

Reference data

Table 4 reports descriptive statistics of each discourse variable (data distribution, means, standard deviations, ranges and medians) for each session. In summary, no significant differences between groups for each dependent variable were revealed. No systematic differences were obtained for both macrostructural and microstructural variables. The strengths of the relationship between test and retest ranged from weak to moderate for all variables.

Test-retest reliability

Test-retest reliability results are presented in Table 5. In summary, ICCs between test and retest ranged from poor to moderate for all variables. Among the macrostructural measures, the highest strength of relationship was found for ICUs/min ($ICC_{[2,1]}$ = 0.695) and TUs/minute ($ICC_{[2,1]}$ = 0.631). For the microstructural variables, the highest strength of relationship, based on Koo and Li (2016) was found for WPM, duration, tokens, CIU_{total} and CIUs/minute.

Bland-Altman plots were created for the microstructural variable that obtained the best and the worst test-retest ICCs. Figure 1 illustrates the limits of agreement for the variables with the highest strengths of relationships, namely ICUs/minute ($ICC_{[2,1]}$ = 0.695), TUs/minute ($ICC_{[2,1]}$ = 0.631) and WPM ($ICC_{[2,1]}$ = 0.641). Mean difference of

Table 3. Frequency for each Information Content Unit.

	#ICU	Description of the ICU	Other acceptable answers	Frequency			
				Test		Retest	
				n	%	n	%
Subjects	Subj1	Homme (qui lit [Action5]) (*man (reading)*)	Monsieur/père/papa/mari/copain/gars/compagnon/ Garçon/Personne qui [Action5] *(mister/dad/daddy/husband/boyfriend/guy/companion/ boy/person who [Action 5])*	63	95	61	92
	Subj2	Homme (qui pêche [Action4]) (*man (fishing)*)	Grand-père/gars/garçon/Personne qui [Action4] ou Personne sur [Place4] *(grandfather/guy/boy/person who [Action4]) or person on [Place4])*	61	92	63	95
	Subj3	Femme *(woman/girl)*	Madame/mère/maman/dame/fille/Personne qui [Action6] (*mrs/mother/mom/lady/girl/person who [Action6]*)	62	94	56	85
	Subj4	Garçon *(boy)*	Enfant/frère/jeune/gars (*kid/brother/young/child*)	63	95	66	100
	Subj5	Fille/enfant *(girl/kid)*	Fillette/jeune/soeur (*little girl/child/sister)*	63	95	65	98
	Subj6	Couple (qui pique-nique [Action1]) *(couple (having a picnic))*	Parents qui [Action1] (*parents (having a picnic)*)	28	42	33	50
	Subj7	Gens/personnes sur le bateau *(people/persons on the boat)*	Quelqu'un sur [Ent6]/plaisanciers (*someone on the [Ent6]/boaters)*	39	59	39	59
	Subj8	Chien *(dog)*		65	98	65	98
Places	Place1	Devant le garage *(in front of the garage)*	Devant la maison/dans l'entrée/dans l'allée (*in front of the house/in the driveway/in the driveway)*	32	48	21	32
	Place2	Sur le bord de l'eau (*on the water's edge*)	Lac/rivière/cours d'eau/mer *(lake/river/watercourse/sea)*	53	80	59	89
	Place3*	Sur la plage/dans le sable *(on the beach/in the sand)*	Sur la terre, rive, rivage, berge, grève *(on the) shore)*	43	65	46	70
	Place4	Sur le quai *(on the dock/on the jetty)*	Sur la jetée	41	62	46	70

(*Continued*)

Table 3. Continued.

	#ICU	Description of the ICU	Other acceptable answers	Frequency			
				Test		Retest	
				n	%	n	%
Entities	Ent1	Cerf-volant *(kite)*		65	98	66	100
	Ent2	Seau/chaudière *(bucket)*		31	47	26	39
	Ent3	Livre/volume *(book)*		20	30	36	39
	Ent4	Breuvage (*drink)*	Boisson/bouteille/quelque chose à boire/verre/vin/ bière/liqueur/de l'eau/*drink/ alcool/liquide* (*beverage/bottle/something to drink/drink/wine/ beer/soft drink/water/drink/ liquid)*	58	88	56	85
	Ent5	Voiture/automobile *(car)*		55	83	47	71
	Ent6	Bateau/voilier *(boat/sailing ship)*		65	98	62	94
	Ent7	Pelle *(spade)*		33	50	28	42
	Ent8	Drapeau *(flag)*		40	61	32	48
	Ent9	Radio *(radio)*	Appareil qui [Action10] (*device that [Action10]*)	52	79	52	79
	Ent10	Panier de pique-nique *(picnic basket)*	Sac/boîte de pique-nique (*picnic bag/box*)	25	38	24	36
	Ent11	Sandales *(sandals)*	Souliers/chaussures (*shoes/footwear*)	33	50	30	45
	Ent12	Arbre *(tree)*		44	67	39	59
	Ent13	Maison *(house)*	Chalet/propriété/demeure (*Cottage/property/home*)	65	98	65	98

(*Continued*)

Table 3. Continued.

	#ICU	Description of the ICU	Other acceptable answers	Frequency			
				Test		Retest	
				n	%	n	%
Actions	Action1	Le couple [Subj6] pique-nique *(Couple having a picnic)*		53	80	58	88
	Action2	Personnes [Subj7] font de la voile *(people sailing)*	[Ent6] vogue, se promène, passe, file [Sub7] font une balade en [Ent6] naviguent/se promènent/voguent en bateau	19	29	19	29
	Action3	Le garçon [Subj4] fait voler le cerf-volant [Ent1] *(boy flying a kite)*	[Sub4] tient/joue/court/s'amuse/traîne/lance/tire/ promène son [Ent1] *([Sub4] holds/plays/runs/has fun/drags/throws/pulls/ walks his [Ent1])*	58	88	54	82
	Action4	L'homme [Subj2] pêche *(man fishing)*	[Subj2] attrape (un poisson/une prise) *([Subj2] catches (a fish))*	58	88	54	82
	Action5	L'homme [Subj1] lit (un livre) *(man reading)*	[Subj1] fait la lecture *([Subj1] reads to)*	56	85	59	89
	Action6	[Subj3] verse (une boisson/à boire) [Subj3] prend (un verre) *(girl pours/has a drink)*	[Subj3] sert/vide (une boisson/à boire) *([Subj3] serves/empties (a drink/to drink))*	55	83	53	80
	Action7	La voiture [Ent5] est stationnée/garée devant le garage [Place1] *(car parked in front of the garage)*	[Ent5] est garée [Place1] *([Ent5] is parked [Place1])*	29	44	23	35
	Action8	[Subj5] joue dans le sable [Place3] *(child playing on the beach)*	[Sub5] construit, bâtit, fabrique, fait (un château de sable) *([Sub5] builds, builds, makes, make (a sandcastle))*	65	98	66	100
	Action9	Drapeau [Ent8] vole *(flag flyes)*	[Ent8] flotte/bouge ([Ent8] *flutters/moves)*	16	24	11	17
	Action10	La radio [Ent9] joue (de la musique) *(radio playing)*	[Ent9] fonctionne, est allumée *([Ent9] is on)* [SubjX] écoute de la musique *([SubjX] listens to music)* Il y a de la musique, on entend de la musique *(There is music, we hear music*)	32	48	31	47
	Action11	Le chien [Subj8] court Le chien [Subj8] poursuit (le garçon) [Subj4] *(dog is running/is chasing)*	[Sub8] gambade/suit [Sub4] *([Sub8] roams/follows [Sub4])* [Sub4] est suivi par le [Sub8] *([Sub4] is followed by [Sub8])*	47	71	54	82

ICU = information content unit; n = number of persons who said the ICU; % = percentage of persons who said the ICU.
*Combination of 'on the beach' and 'in the sand' from the list used by Jensen et al. (2006)'s as they were used interchangeably in French.

Table 4. Descriptive statistics of macrotructural and microstructural variables of discourse and statistics comparing group difference between two sessions (T1 and T2).

	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)	
Macrostructural variables							
ICU_{total}	25.56 (4.61)	26 [14–33]	25.076 (4.916)	25 [14–33]	768 (p =.38)	0.49 (p <.001)	No systematic difference, moderate relationship between sessions.
ICUs per minute	21.62 (8.73)	20.67 [8.31–48.89]	20.310 (7.454)	18.61 [6.33–41.67]	847 (p =.10)	0.65 (p <.001)	No systematic difference, moderate relationship between sessions.
ICUs/utterance	1.36 (0.49)	1.34 [0.43–3.14]	1.292 (0.465)	[0.31–2.57]	940 (p =.39)	0.53 (p <.001)	No systematic difference, moderate relationship between sessions.
$ICU_{subjects}$	6.73 (1.20)	7 [1–8]	6.788 (0.953)	7 [5–8]	481 (p =.92)	0.30 (p =.013)	No systematic difference, weak relationship between sessions.
ICU_{places}	2.56 (1.05)	3 [0–4]	2.606 (1.006)	3 [0–4]	418 (p =.68)	0.39 (p <.001)	No systematic difference, weak relationship between sessions.
$ICU_{entities}$	8.88 (2.61)	9 [3–13]	8.379 (3.012)	8 [2–13]	689 (p =.20)	0.52 (p <.001)	No systematic difference, moderate relationship between sessions.
$ICU_{actions}$	7.39 (1.51)	8 [4–11]	7.303 (1.598)	7 [3–10]	707.50 (p =.76)	0.23 (p =.061)	No systematic difference, weak relationship between sessions.
TU_{total}	15.17 (1.296)	16 [11–16]	15.02 (1.271)	15 [11–16]	289 (p =.02†)	0.36 (p =.003)	No systematic difference after correction for multiple comparisons, weak relationship between sessions.
TUs per minute	13.29 (6.24)	12.08 [4.75–35.56]	12.619 (5.262)	11.84 [3.07–27.10]	966 (p =.37)	0.58 (p <.001)	No systematic difference, moderate relationship between sessions.
TUs per utterance	0.83 (0.31)	0.81 [0.25–1.88]	0.803 (0.333)	0.74 [0.15–1.71]	1001 (p =.64)	0.49 (p <.001)	No systematic difference, moderate relationship between sessions.
Microstructural variables							
Duration (seconds)	84.12 (40.88)	77 [26–202]	85.94 (44.83)	77 [31–313]	1115 (p =.78)	0.62 (p <.001)	No systematic difference, moderate relationship between sessions.
Tokens	230.97 (115.43)	208.50 [74–661]	239.73 (128.58)	217.50 [86–820]	1164 (p =.71)	0.64 (p <.001)	No systematic difference, moderate relationship between sessions.
MLU (words)	10.21 (2.20)	9.91 [6.36–15.20]	10.16 (2.07)	9.73 [6.55–15.14]	1096 (p =.95)	0.40 (p <.001)	No systematic difference, moderate relationship between sessions.
Propositionnal density	0.36 (0.05)	0.36 [0.25–0.50]	0.37 (0.05)	0.37 [0.29–0.47]	1307 (p =.47)	0.41 (p =.001)	No systematic difference, moderate relationship between sessions.

(*Continued*)

Table 4. Continued.

	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)	
Words per minute	167.39 (29.66)	164.33 [107.30–251.10]	168.54 (25.76)	168.03 [107.08–236.32]	1140 (p =.83)	0.58 (p <.001)	No systematic difference, moderate relationship between sessions.
Verbs per utterance	0.53 (0.23)	0.49 [0.10–1.30]	0.50 (0.197)	0.48 [0.07–1.09]	994 (p =.48)	0.38 (p =.002)	No systematic difference, weak relationship between sessions.
Open/closed ratio	1.03 (0.12)	1.02 [0.82–1.37]	1.06 (0.14)	1.05 [0.81–1.56]	1441 (p =.03†)	0.43 (p <.001)	No systematic difference after correction for multiple comparisons, moderate relationship between sessions.
Noun-to-verb ratio	6.36 (3.75)	6.06 [1.77–30.00]	6.94 (5.798)	5.64 [2.48–37.00]	1006 (p =.52)	0.45 (p <.001)	No systematic difference, moderate relationship between sessions.
MATTR	0.95 (0.01)	0.95 [0.91–0.98]	0.95 (0.02)	0.96 [0.89–0.99]	1158 (p =.58)	0.42 (p =.050)	No systematic difference, moderate relationship between sessions.
CIU_{total}	223.27 (109.62)	194.50 [77–637]	235.64 (124.83)	213.50 [85–802]	1182 (p =.47)	0.62 (p <.001)	No systematic difference, moderate relationship between sessions.
Percentage of CIUs	95.39 (3.67)	96.68 [85.15–100]	94.62 (3.58)	95.23 [82.49-100]	913 (p =.22)	0.31 (p =.012)	No systematic difference, weak relationship between sessions.
CIUs per minute	164.26 (37.63)	158.62 [70.84–319.41]	166.34 (26.71)	164.34 [112.31–233.85]	1186 (p =.61)	0.63 (p <.001)	No systematic difference, moderate relationship between sessions.

†Non-significant when correcting for multiple comparisons using the Bonferroni correction.
SD = standard deviation; min = minimum; max = maximum; ICU = information content unit; TU = thematic unit; MLU = mean length of utterance; MATTR = moving-average type-token ratio; CIU = correct information unit.

Table 5. Summary of test-retest results.&

Measure	ICC			Correlation		Absolute Value Difference Between Test and Retest		MDC90
	ICC	95% CI Low – High	Koo and Li (2016) ICC Quality [CI Quality]	Spearman' rho	p value	M (SD)	Range	
Macrostructural variables								
ICU_{total}	0.535	0.338–0.687	Moderate [Poor-Moderate]	0.49	< 0.001	3.70 (2.74)	0.00–13.00	5.25
ICUs per minute	0.695	0.546–0.801	Moderate [Moderate-Good]	0.65	< 0.001	5.07 (3.88)	0.22–16.31	8.97
ICUs per utterance	0.544	0.351–0.693	Moderate [Poor-Moderate]	0.53	< 0.001	5.07 (3.88)	0.22–16.31	0.53
$ICU_{subjects}$	0.347	0.115–0.543	Poor [Poor-Moderate]	0.30	0.013	0.91 (0.84)	0–4	1.19
ICU_{places}	0.446	0.229–0.621	Poor [Poor-Moderate]	0.39	0.001	0.77 (0.76)	0–3	1.14
$ICU_{entities}$	0.504	0.303–0.663	Moderate [Poor-Moderate]	0.52	< 0.001	2.20 (1.77)	0–8	3.12
$ICU_{actions}$	0.316	0.080–0.518	Poor [Poor-Moderate]	0.23	0.061	1.39 (1.61)	0–5	1.71
TU_{total}	0.373	0.146–0.563	Poor [Poor-Moderate]	0.36	0.003	0.93 (1.09)	0–5	1.42
TUs per minute	0.631	0.461–0.756	Moderate [Poor-Good]	0.58	< 0.001	3.85 (3.15)	0.20–12.82	6.37
TUs per utterance	0.488	0.280–0.652	Poor [Poor-Moderate]	0.49	< 0.001	(0.21)	0.00–1.00	0.35
Microstructural variables								
Duration (seconds)	0.601	0.421–0.736	Moderate [Poor-Moderate]	0.62	< 0.001	27.79 (26.44)	0–124	47.32
Tokens	0.580	0.395–0.720	Moderate [Poor-Moderate]	0.64	< 0.001	78.09 (80.48)	0.06–1.88	134.81
MLU (words)	0.393	0.166–0.579	Poor [Poor-Moderate]	0.40	< 0.001	1.88 (1.41)	0.06–6.48	2.36
Propositionnal density	0.452	0.237–0.625	Poor [Poor-Moderate]	0.41	0.001	0.04 (0.03)	0.00–0.14	0.05
Words per minute	0.641	0.473–0.764	Moderate [Poor-Good]	0.58	< 0.001	18.62 (14.43)	0.23–69.98	30.64
Verbs per utterance	0.408	0.187–0.590	Poor [Poor-Moderate]	0.38	0.002	0.19 (0.14)	0.01–0.75	0.24
Open/closed ratio	0.393	0.174–0.577	Poor [Poor-Moderate]	0.43	< 0.001	0.12 (0.09)	0.00–0.45	0.15
Noun-to-verb ratio	0.265	0.025–0.475	Poor [Poor]	0.45	< 0.001	0.19 (0.14)	0.01–0.75	5.40
MATTR	0.244	0.004–0.458	Poor [Poor-Moderate]	0.42	0.050	0.01 (0.01)	0.00–0.05	0.02
CIU_{total}	0.575	0.389–0.716	Moderate [Poor-Moderate]	0.62	< 0.001	76.48 (77.24)	0–371	129.70
Percentage of CIUs	0.420	0.204–0.599	Poor [Poor-Moderate]	0.31	0.012	3.03 (3.52)	0.02–9.40	4.02
CIUs per minute	0.543	0.348–0.694	Moderate [Poor-Moderate]	0.63	< 0.001	22.13 (22.00)	0.13–128.16	35.99

$n = 66$.

*not significant using the adjusted *p*value following the Bonferroni correction ($p < .005$).

SD = Standard Deviation; CI = Confidence Interval; MC_{total} = Main Concept total score; AC = Accurate and Complete; AI = Accurate and Incomplete; IC = Incorrect and Complete; II = Incorrect and Incomplete; AB = Absent; MLU = Mean Length of Utterances; CIU = Correct Information Units; MATTR = Moving-Average Type-Token Ratio; MDC90= Minimal Detectable Change at 90% confidence.

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.

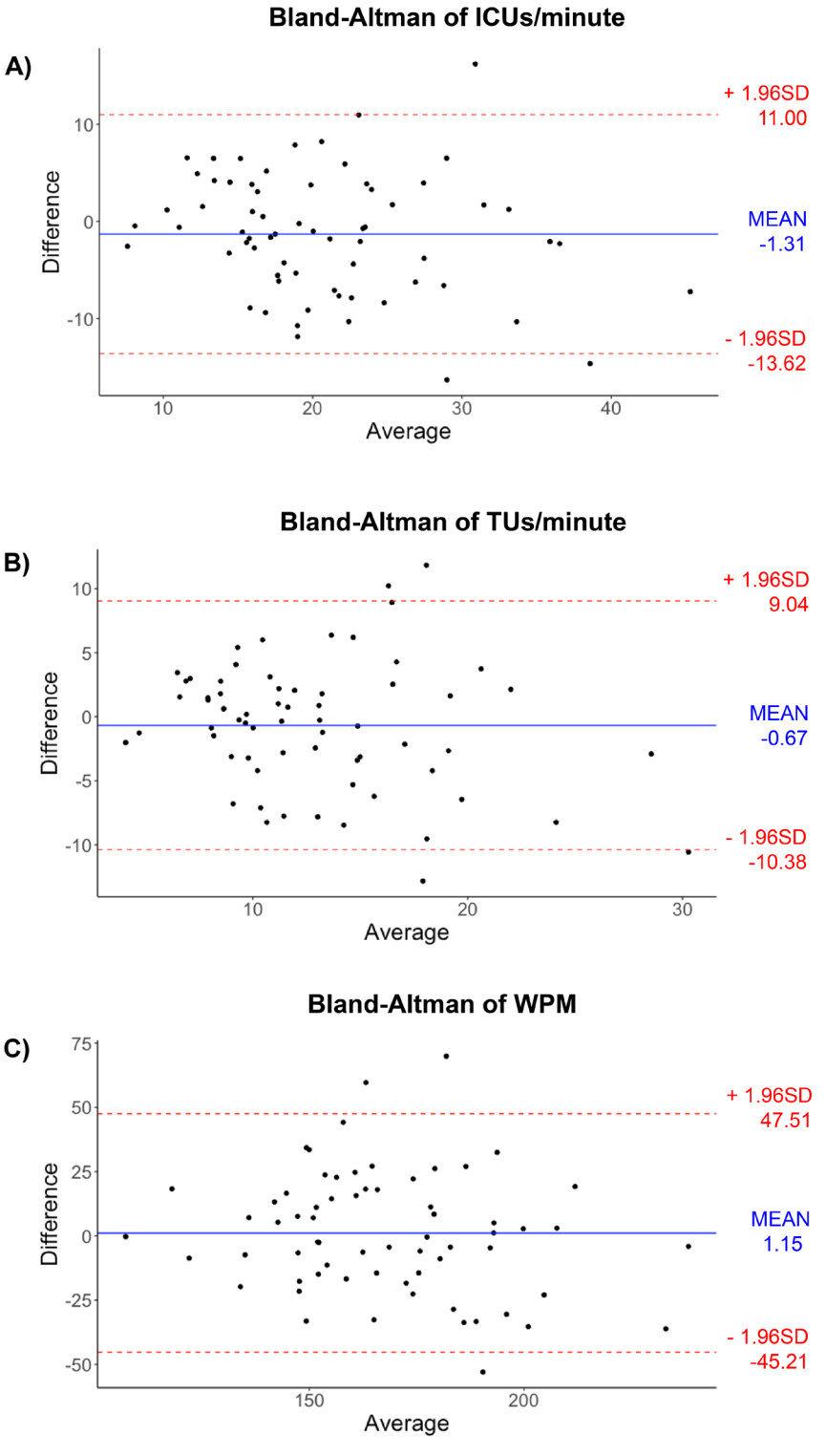

Figure 1. Bland-Altman plots for the variables with the highest strengths of relationships. The upper plot (a) represents the limits of agreement for ICUs/minute, the middle plot (b) represents TUs/minute and the lower plot (c) represents WPM.

Legend: ICU = information content unit; TU = thematic unit; WPM = words per minute; SD = standard deviation

agreement between test and retest was the closest to zero for TUs/minute, more precisely at −0.67. However, only ICUs/minute and WPM demonstrated good agreement according to the standards of Bland and Altman (1999), with 95% of the data (i.e. 63 out of 66) within ±1.96 standard deviations of the mean of differences. TUs/minute obtained 90% of the values (i.e. 62 out of 66) within limits of agreement of ±1.96 standard deviations.

Figure 2 represents the limits of agreement for the variables with the lowest strengths of relationships, namely noun-to-verb ratio ($ICC_{[2,1]}$ = 0.265) and MATTR ($ICC_{[2,1]}$ = 0.244). Although the strengths of the relationships were poor, both

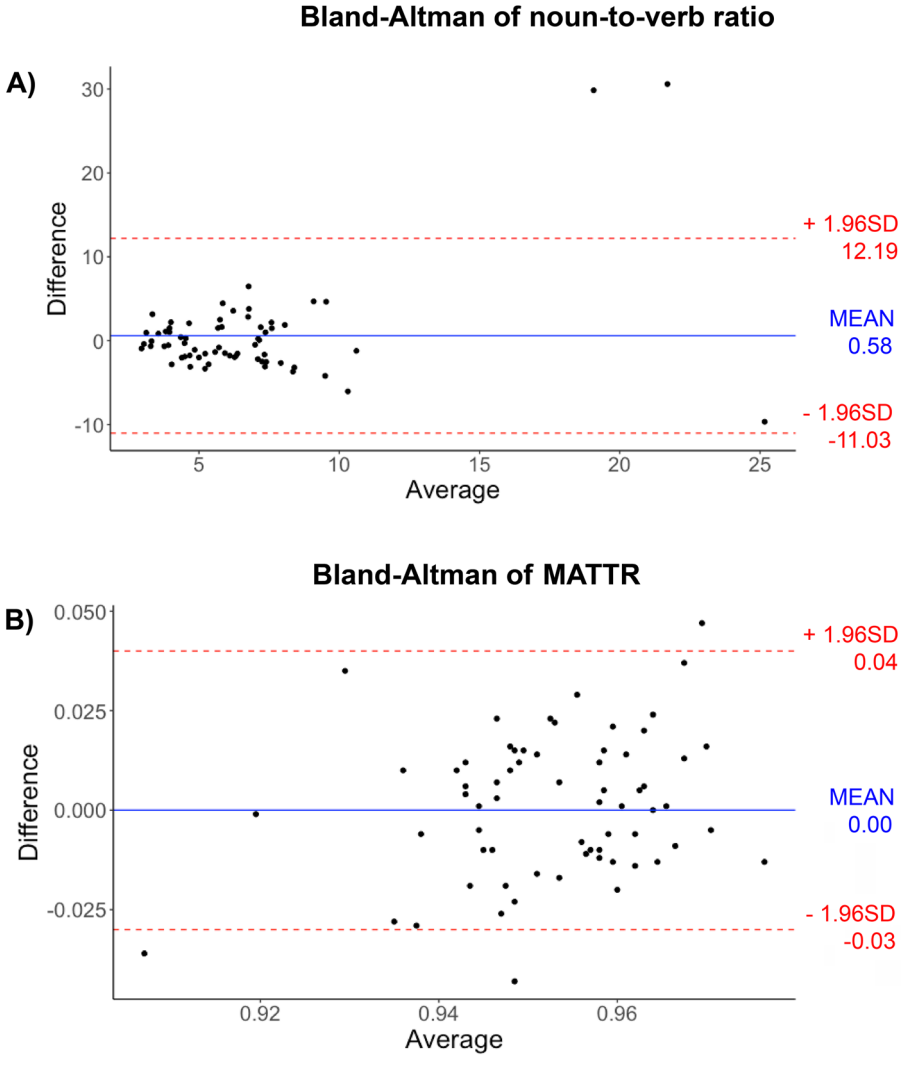

Figure 2. Bland-Altman plots for the variables with the lowest strengths of relationships. The upper plot (a) represents the limits of agreement for noun-to-verb ratio and the lower plot (b) represents MATTR.
Legend: MATTR = moving average type/token ratio; SD = standard deviation

noun-to-verb ratio and MATTR demonstrated good agreement according to the standards of Bland and Altman (1999), with 95% of the data (i.e. respectively 64 and 63 out of 66) within ±1.96 standard deviations of the mean of differences. The mean difference of agreement between test and retest was of zero for MATTR and close to zero for noun-to-verb ratio (0.58).

Discussion

This study aimed to document inter-rater and test-retest reliability of various discourse measures in Laurentian French, including the cultural adaptation of an ICU list, and to provide reference data for the picture description task of the picnic scene (Kertesz, 2006) in PWBI. Firstly, a cultural and linguistic adaptation of the ICU list of Jensen et al. (2006) was developed to reflect speakers of Laurentian French. Similar to our adaptation of the main concept analysis for the Cinderella story retell task (Brisebois et al., 2023), our adaptation of the ICU list task led to modifications from the original list. Regarding reliability, inter-rater reliability results ranged from good to excellent for all variables. While there were no systematic differences between test and retest for all variables, test-retest reliability was poor to moderate. As a result, used alone, this discourse task does not meet the requirements to conduct group research studies in PWBI (ICC >.70), and even less for clinical use (ICC >.90) (Fitzpatrick et al., 1998).

Test-retest reliability

The results of the current study are complementary to the previous studies conducted by members of our research team (Boucher et al., 2022; Brisebois et al., 2023; Marcotte et al., 2022) that aimed to develop gold standard measures to assess discourse production in PWBI who speak Laurentian French. These results highlighted the continued need to investigate test-retest reliability of discourse measures (Pritchard et al., 2017, 2018). The lack of valid and standardized discourse measures compromises the early detection of pathological changes and does not allow clinicians to fully capture the changes between two assessments. Not surprisingly, the most reliable discourse measures were those of efficiency, namely ICUs/min, TUs/min and WPM. This is consistent with previous findings (Boyle, 2015; Brookshire & Nicholas, 1994; Stark et al., 2023), which reported that WPM and CIUs/min are reliable measures to use for clinical decision making in people with aphasia as well as for the detection of subtle or mild cognitive decline. Consistent with previous evidence, this study suggests that WPM and ICUs/min are among the most reliable measures in PWBI.

Nonetheless, in contrast to our recent work with the Cinderella story retell task (Brisebois et al., 2023), no measures extracted from the picnic scene met the reliability requirements as defined by Boyle (2014) and Fitzpatrick et al. (1998) for inclusion in research studies and, even less, the criterion for clinical use. Considering the poor test-retest reliability of this task, we recommend selecting the measures with the highest test-retest reliability, and to use them with caution in both research and clinical making decisions.

We did not compare test-retest reliability between PWBI and people with aphasia, but previous evidence suggests that it is generally lower in PWBI (Brookshire & Nicholas, 1994; Stark et al., 2023). Therefore, the normative data presented here should not be used to evaluate the recovery or the impact of therapy for people with aphasia. Further research will be needed to establish the psychometric properties of this discourse task in a group of people with chronic aphasia, considering the differences observed between the two groups by others (Stark et al., 2023). Test-retest reliability of discourse measures improves when looking at a set of tasks rather than when evaluating for each task separately (Boyle, 2014; Brookshire & Nicholas, 1994; Stark et al., 2023). As recently highlighted by Stark et al. (2023), it is crucial to evaluate the test-retest reliability of each task or each set of tasks because of the variability observed between the different tasks.

Inter-rater reliability

Inter-rater reliability (IRR) is also an important psychometric property to consider when trying to identify outcome measures. As reported previously (Boucher et al., 2022; Marcotte et al., 2022), the total number of TUs and total number of ICUs in our study showed excellent IRR. Consistent with previous studies, including ours (e.g. Brisebois et al., 2023; Stark et al., 2023), CIUs and tokens also produced excellent IRR. The excellent reliability for the tokens suggests that the transcriptions were highly reliable between our raters. In contrast, IRR for utterances (i.e. which refers to the segmentation of the sample into utterances) was only considered good in the present study, but excellent in previous studies including a recent study by our group (e.g. Brisebois et al., 2023; Stark et al., 2023). Although IRR was found to be excellent with utterances using the Cinderella story retell task in our recent study, it was lower than in Stark et al. (2023), which may be explained by three main differences. First, the prosody in French is very different than that in English. Briefly, French is characterized by a succession of mostly rising contours for non-terminal constituents, and a greater variability for the tonal contour of terminal constituents (Delais-Roussarie et al., 2020). Thus, the lower regularity in prosody of the terminal constituents in French may confound segmentation. Colin and Le Meur (2016) added supplementary rules to reduce the difficulty associated with segmentation in French, but they still reported that it was difficult to standardize the segmentation. Second, the length of this study's samples was shorter than in previous studies, which either combined the discourse tasks (Stark et al., 2023) or had longer samples (Brisebois et al., 2023). The lower number of utterances might have reduced the statistical power of the ICC. Samples of a minimum of 300 to 400 are recommended to improve test-retest reliability (Brookshire & Nicholas, 1994). In the current study, we collected samples with a mean of 234 words at test and 250 words at retest, which is below the recommended minimum length of samples that investigate test-retest reliability. Third, it is now well documented that the instructions given and the pictorial stimulus used to elicit discourse generate differences in the style of production. For instance, picture description tasks such as the picnic scene reduce the use of linguistic markers to connect the different elements, which may complicate the segmentation compared to other types of tasks (Marini et al., 2005).

Clinical implications

One of the main applications of our study is the elaboration of a list of ICUs in Laurentian French for the picnic scene from the WAB-R. In their review, Slegers et al. (2018) showed that information units and efficiency were the most reported discriminant variables in picture description tasks between individuals with AD and PWBI. Documentation of macro-structural features in discourse in healthy adults over time is interesting because studies suggest that changes in measures relating to the macrostructure of discourse (i.e. informativeness, global and local coherence) may elicit deficits associated with the decline of cognitive functions in neurocognitive disorders (Pistono et al., 2019; Slegers et al., 2018; Taler & Phillips, 2008). Another important reason for the adaptation of the ICU list for the picnic scene to Laurentian French is that this measure is relatively easy and quick to implement in language assessments, including for both PWBI and people with aphasia. Microstructural analyses typically rely on long transcriptions which are used less frequently in clinical settings (Bryant et al., 2017). Similar to our TU list (Brisebois et al., 2020), the ICU scoring list is based on a finite set of content units that are more easily quantified and thus more suitable for clinical settings.

Moreover, the present study provides reference data regarding the longitudinal changes in discourse of PWBI. We reported variability and Minimal Detectable Change (MDC90), which are essential in both clinical settings and future studies to identify "real" changes and not only changes associated with normal test-retest variability, especially in subclinical populations. For instance, considering that SCD is usually not detected by standard cognitive testing, its identification requires measures highly sensitive and with robust psychometrical features (Jessen et al., 2014). In literature reviews of discourse measures in people with neurocognitive diseases (e.g. Filiou et al., 2020; Slegers et al., 2018), microstructural variables were identified to be different in people with mild cognitive impairment compared to PWBI in picture description tasks (Filiou et al., 2020; Slegers et al., 2018). However, limited data are available regarding the normal variability observed between two testing sessions. The adaptation and characterization of the reliability of discourse measures for Laurentian French speakers is thus potentially important for clinicians to profile impairments associated with neurocognitive conditions (Croisile et al., 1996; Gallée et al., 2021; Jensen et al., 2006), or SCD. As mentioned previously, a large proportion of speech-language pathologists (Bryant et al., 2017), and probably neuropsychologists, only use one discourse task in their assessments. The present results suggest that both researchers and clinicians should be careful in their interpretation of change with the description of the picnic scene of the WAB-R (Kertesz, 2006) when used alone.

The importance of increasing linguistic and cultural diversity

An urgent global call for action was recently made by the International Network for Cross-Linguistic Research on Brain Health, better known as Include (https://include-network.com), to increase linguistic and cultural diversity in the investigation of neurocognitive disorders (García et al., 2023). To date, the majority of studies have been conducted with English speakers. The present study aims to help reduce the

global inequities across minority languages, by collecting data in an under-represented language such as Laurentian French. By doing so, we have contributed to the generation of linguistic features that are potentially able to differentiate between normal aging, SCD and various neurocognitive diseases, by using cost-effective language assessments and by developing rigorous and standardized discourse measures. An increased number of studies on languages other than English is critical to reduce global inequities concerning the assessment of neurocognitive diseases.

Limitations

This study is not without limitations. First, the sample size is relatively small, although comparable to e.g. Richardson and Dalton (2016) or even higher than similar studies (Stark et al., 2023). Considering that the population studied (i.e. French-speaking persons living in Quebec) is less than 8 million people, the number of participants is relatively high compared to similar studies. Second, the present results might not be generalizable to other French dialects because some words and expressions in Laurentian French are only used in this specific dialect. Third, our sample lacks representation of individuals with lower levels of education. All participants had a minimum of 11 years of education (i.e. high school completed in Quebec). Previous evidence has demonstrated the impact of education on discourse abilities. For instance, people with fewer years of education tend to produce shorter and incomplete descriptions (Mackenzie, 2000). Similarly, Le Dorze and Bédard (1998) reported that Laurentian French speakers with fewer years of education produced less informative discourse. It will be important in the future to include individuals with lower levels of education. Fourth, in contrast to previous studies (e.g. Stark et al., 2023), the time between testing sessions was longer and ranged from 162 to 406 days, to better reflect changes associated with typical aging (Mueller et al., 2018) and the time between two assessments when neurocognitive disease is suspected. This made comparison with other studies difficult. Fifth, a cognitive screening was not administered to all our participants (and has therefore not been reported). However, no participant self-reported any cognitive impairments nor any impact on their daily functioning. Sixth, no vision nor auditory screenings were conducted to ensure all participants had sufficient vision and hearing abilities.

Conclusion

To conclude, we have developed a linguistically and culturally adapted ICU list and documented poor to moderate test-retest reliability of discourse measures in speakers of Laurentian French without brain injury for the picnic scene of the WAB-R (Kertesz, 2006). The present study contributes to the urgent need to increase linguistic and cultural diversity in the investigation of spoken discourse and provide tools for early detection of neurocognitive disorders (García et al., 2023). It is also crucial to be able to detect the presence of pathological changes in PWBI. The scarcity of psychometrically robust normative data for Laurentian French, a non-dominant language in North America, creates inequities for this minority population and is a barrier to assessing discourse production for both researchers and clinicians.

The picnic scene is used by several clinicians and researchers who work with speakers of Laurentian French, just as it is to Canadian speakers of English, because it illustrates a typical scene commonly experienced (or observed) in Quebec. Thus, the cultural adaptation of the ICU list of the picnic scene is important. The overall results provide insight into typical performance and variation, which is crucial to differentiate language changes due to pathology (Boyle, 2014). Considering the multitude of factors that can have an impact on intra-individual variability and test-retest reliability, this study supports the refinement of the psychometric properties of measures available for discourse analysis for Laurentian French speakers in Quebec.

Acknowledgments

We are very grateful to all the participants who contributed to this study. Also, we wish to thank Marianne Désilets-Barnabé and Noémie Desjardins for their help in data collection. 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. S.M.B. K.M. holds a Career Award from the "Fonds de Recherche du Québec – Santé." A.B. holds a scholarship from the "Fonds de Recherche du Québec – Santé."

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

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. holds a Career Award from the "Fonds de Recherche du Québec – Santé." A.B. holds a scholarship from the "Fonds de Recherche du Québec – S.

Data availability statement

The raw data presented in this article are not readily available because of the sensitivity of the video materials. The datasets analysed for the current study are available from the corresponding author upon reasonable request.

References

Ahmed, S., Haigh, A.-M F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain: A Journal of Neurology*, *136*(Pt 12), 3727–3737. https://doi.org/10.1093/brain/awt269

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (the Statistician)*, *32*(3), 307–317. https://doi.org/10.2307/2987937

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

Azios, J. H., Archer, B., Simmons-Mackie, N., Raymer, A., Carragher, M., Shashikanth, S., & Gulick, E. (2022). Conversation as an outcome of aphasia treatment: A systematic scoping review.

American Journal of Speech-Language Pathology, *31*(6), 2920–2942. https://doi.org/10.1044/2022_AJSLP-22-00011

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, *11*(1), 118. https://doi.org/10.1186/1471-2202-11-118

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204open_in_new

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*, *8*(, 269. https://doi.org/10.3389/fpsyg.2017.00269

Boucher, J., Brisebois, A., Slegers, A., Courson, M., Désilets-Barnabé, M., Chouinard, A.-M., Gbeglo, V., Marcotte, K., & Brambati, S. M. (2022). Picture description of the western aphasia battery picnic scene: Reference data for the French Canadian Population. *American Journal of Speech-Language Pathology*, *31*(1), 257–270. https://doi.org/10.1044/2021_AJSLP-20-00388

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

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 neurotypical speakers. *American Journal of Speech-Language Pathology*, *32*(6), 2871–2888. https://doi.org/10.1044/2023_AJSLP-23-00101

Brisebois, A., Brambati, S. M., Rochon, E., Leonard, C., & Marcotte, K. (2023). The longitudinal trajectory of discourse from the hyperacute to the chronic phase in mild to moderate post-stroke aphasia recovery: A case series study. *International Journal of Language & Communication Disorders*, *58*(4), 1061–1081. https://doi.org/10.1111/1460-6984.12844

Brookshire, R. H., & Nicholas, L. E. (1984). Comprehension of directly and indirectly stated main ideas and details in discourse by brain-damaged and non-brain-damaged listeners. *Brain and Language*, *21*(1), 21–36. - 36 https://doi.org/10.1016/0093-934X(84)90033-6

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

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

Callahan, B. L., Macoir, J., Hudon, C., Bier, N., Chouinard, N., Cossette-Harvey, M., Daigle, N., Fradette, C., Gagnon, L., & Potvin, O. (2010). Normative data for the pyramids and palm trees test in the Quebec-French population. *Archives of Clinical Neuropsychology*, *25*(3), 212–217. https://doi.org/10.1093/arclin/acq013

Capilouto, G. J., Wright, H. H., & Maddy, K. M. (2016). Microlinguistic processes that contribute to the ability to relay main events: Influence of age. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *23*(4), 445–463. https://doi.org/10.1080/13825585.2015.1118006

Chenery, H. J., & Murdoch, B. E. (1994). The production of narrative discourse in response to animations in persons with dementia of the Alzheimer's type: Preliminary findings. *Aphasiology*, *8*(2), 159–171. https://doi.org/10.1080/02687039408248648

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 [French adaptation of the AphasiaBank project]* [Univeristé Paul Sabatier]. https://aphasia.talkbank.org/access/French/0docs/ColinLeMeurmemoire.pdf

Cooper, P. V. (1990). Discourse production and normal aging: Performance on oral picture description tasks. *Journal of Gerontology*, *45*(5), P210–P214. https://doi.org/10.1093/geronj/45.5.P210

Criel, Y., Deleu, M., De Groote, E., Bockstael, A., Kong, A. P.-H., & De Letter, M. (2021). The Dutch main concept analysis: Translation and establishment of normative data. *American Journal of Speech-Language Pathology*, *30*(4), 1750–1766. https://doi.org/10.1044/2021_AJSLP-20-00285

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language*, *53*(1), 1–19. https://doi.org/10.1006/brln.1996.0033

DeepL Traduction—DeepL Translate. (2022). Le meilleur traducteur au monde. https://www.DeepL.com/translator

Delais-Roussarie, E., Post, B., & Yoo, H. (2020). Prosodic units and intonational grammar in French: Towards a new approach. *Speech Prosody*, *2020*, 126–130. https://doi.org/10.21437/SpeechProsody.2020-26

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? *Aphasiology*, *32*(4), 459–464. https://doi.org/10.1080/02687038.2017.1398803

Dijk, T. A. V. (2019). *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition* (1re éd.). Routledge. https://doi.org/10.4324/9780429025532

Donoghue, D., & Stokes, E. K, Physiotherapy Research and Older People (PROP) group. (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

Doyle, P. J., Goda, A. J., & Spencer, K. A. (1995). The Communicative informativeness and efficiency of connected discourse by adults with aphasia under structured and conversational sampling conditions. *American Journal of Speech-Language Pathology*, *4*(4), 130–134. https://doi.org/10.1044/1058-0360.0404.130

Duong, A., Tardif, A., & Ska, B. (2003). Discourse about discourse: What is it and how does it progress in Alzheimer's disease? *Brain and Cognition*, *53*(2), 177–180. https://doi.org/10.1016/S0278-2626(03)00104-0

Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H. (2019). Modeling confrontation naming and discourse informativeness using structural equation modeling. *Aphasiology*, *33*(5), 544–560. https://doi.org/10.1080/02687038.2018.1482404

Filiou, R.-P., Bier, N., Slegers, A., Houzé, B., Belchior, P., & Brambati, S. M. (2020). Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: A scoping review. *Aphasiology*, *34*(6), 723–755. https://doi.org/10.1080/02687038.2019.1608502

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), i-iv, 1-74. https://doi.org/10.3310/hta2140

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

Gallée, J., Cordella, C., Fedorenko, E., Hochberg, D., Touroutoglou, A., Quimby, M., & Dickerson, B. C. (2021). Breakdowns in informativeness of naturalistic speech production in primary progressive aphasia. *Brain Sciences*, *11*(2), 130. https://doi.org/10.3390/brainsci11020130

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: A Journal of Neurology*, *146*(12), 4870–4879. https://doi.org/10.1093/brain/awad253

Giles, E., Patterson, K., & Hodges, J. R. (1996). Performance on the Boston Cookie theft picture description task in patients with early dementia of the Alzheimer's type: Missing information. *Aphasiology*, *10*(4), 395–408. https://doi.org/10.1080/02687039608248419

Goodglass, H., Kaplan, E., & Baresi, B. (2001). Boston *diagnostic aphasia examination*. –Third Edition (BDAE-3) Lippincott Williams & Wilkins.

Hillis, A. E. (2007). Aphasia: Progress in the last quarter of a century. *Neurology*, *69*(2), 200–213. https://doi.org/10.1212/01.wnl.0000265600.69385.6f

Jensen, A. M., Chenery, H. J., & Copland, D. A. (2006). A comparison of picture description abilities in individuals with vascular subcortical lesions and Huntington's Disease. *Journal of Communication Disorders*, *39*(1), 62–77. https://doi.org/10.1016/j.jcomdis.2005.07.001

Jessen, F., Amariglio, R. E., Buckley, R. F., Flier, W. M. V D., Han, Y., Molinuevo, J. L., Rabin, L., Rentz, D. M., Rodriguez-Gomez, O., Saykin, A. J., Sikkes, S. A. M., Smart, C. M., Wolfsgruber, S., & Wagner, M. (2020). The characterisation of subjective cognitive decline. *The Lancet Neurology*, *19*(3), 271–278. https://doi.org/10.1016/S1474-4422(19)30368-0

Jessen, F., Amariglio, R. E., van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois, B., Dufouil, C., Ellis, K. A., van der Flier, W. M., Glodzik, L., van Harten, A. C., de Leon, M. J., McHugh, P., Mielke, M. M., Molinuevo, J. L., Mosconi, L., Osorio, R. S., Perrotin, A., … Wagner, M, Subjective Cognitive Decline Initiative (SCD-I) Working Group. (2014). A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's & Dementia*, *10*(6), 844–852. https://doi.org/10.1016/j.jalz.2014.01.001

Kertesz, A. (2006). *Western aphasia battery – Revised*. Pearson.

Kong, A. P.-H. (2009). The use of main concept analysis to measure discourse production in Cantonese-speaking persons with aphasia: A preliminary report. *Journal of Communication Disorders*, *42*(6), 442–464. https://doi.org/10.1016/j.jcomdis.2009.06.002

Kong, A. P.-H. (2016). *Analysis of neurogenic disordered discourse production* (0 éd.). Routledge. https://doi.org/10.4324/9781315639376

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

Le Dorze, G., & Bédard, C. (1998). Effects of age and education on the lexico-semantic content of connected speech in adults. *Journal of Communication Disorders*, *31*(1), 53–71. https://doi.org/10.1016/S0021-9924(97)00051-8

Linnik, 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

Mackenzie, C. (2000). The relevance of education and age in the assessment of discourse comprehension. *Clinical Linguistics & Phonetics*, *14*(2), 151–161. https://doi.org/10.1080/026992000298887

Mackenzie, C., Brady, M., Norrie, J., & Poedjianto, N. (2007). Picture description in neurologically normal adults: Concepts and topic coherence. *Aphasiology*, *21*(3-4), 340–354. https://doi.org/10.1080/02687030600911419

MacWhinney, B. (2000). *The CHILDES project: Tolls for analzying talk* (3rd ed.) Lawrence Erlbaum Associates.

MacWhinney, B., Fromm, D., Forbes, M., & Audrey Holland (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, *25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893

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

Marini, A., Andreetta, S., del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology*, *25*(11), 1372–1392. https://doi.org/10.1080/02687038.2011.584690

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

Mehler, J. (1994). Cross-linguistic approaches to speech processing. *Current Opinion in Neurobiology*, *4*(2), 171–176. https://doi.org/10.1016/0959-4388(94)90068-X

Mitchell, A. J., Beaumont, H., Ferguson, D., Yadegarfar, M., & Stubbs, B. (2014). Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: Meta-analysis. *Acta Psychiatrica Scandinavica*, *130*(6), 439–451. https://doi.org/10.1111/acps.12336

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. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research*, *38*(1), 145–156. https://doi.org/10.1044/jshr.3801.145

Pistono, A., Pariente, J., Bézy, C., Lemesle, B., Le Men, J., & Jucla, M. (2019). What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia*, *124*, 133–143. https://doi.org/10.1016/j.neuropsychologia.2018.12.018

Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, 18 (12), 1075-1091. https://doi.org/10.1080/02687030444000534

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

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), 1078–1093. https://doi.org/10.1111/1460-6984.12420

Richardson, J. D., & Dalton, S. G. H. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. *Aphasiology*, *30*(1), 45–73. https://doi.org/10.1080/02687038.2015.1057891

Schiavetti, N. E., Metz, D. E., & Orlikoff, R. F. (2011). *Evaluation research in communication disorders* (6th ed.). Pearson Education.

Simmons-Mackie, N., & Lynch, K. E. (2013). Qualitative research in aphasia: A review of the literature. *Aphasiology*, *27*(11), 1281–1301. https://doi.org/10.1080/02687038.2013.818098

Ska, B., Duong, A., & Joanette, Y. (2004). Discourse impairments. In R. D. Kent (Éd.), *The MIT encyclopedia of communication disorders* (pp. 302–304). The MIT Press.

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: JAD*, *65*(2), 519–542. https://doi.org/10.3233/JAD-170881

Sloetjes, H., & Wittenburg, P. (2008). *Annotation by category-ELAN and ISO DCR*.

Slot, R. E. R., Sikkes, S. A. M., Berkhof, J., Brodaty, H., Buckley, R., Cavedo, E., Dardiotis, E., Guillo-Benarous, F., Hampel, H., Kochan, N. A., Lista, S., Luck, T., Maruff, P., Molinuevo, J. L., Kornhuber, J., Reisberg, B., Riedel-Heller, S. G., Risacher, S. L., Roehr, S., … van der Flier, W. M, SCD-I working group. (2019). Subjective cognitive decline and rates of incident Alzheimer's disease and non–Alzheimer's disease dementia. *Alzheimer's & Dementia*, *15*(3), 465–476. https://doi.org/10.1016/j.jalz.2018.10.003

Spencer, E., Bryant, L., & Colyvas, K. (2020). Minimizing Variability in language sampling analysis: A practical way to calculate text length and time variability and measure reliable change when assessing clients. *Topics in Language Disorders*, *40*(2), 166–181. https://doi.org/10.1097/TLD.0000000000000212

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*, *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*, *37*(5), 761–784. https://doi.org/10.1080/02687038.2022.2039372

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., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., Ramage, A. E., & Roberts, A. C. (2021). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *Journal of Speech, Language, and Hearing Research*, *64*(11), 4366–4389. https://doi.org/10.1044/2021_JSLHR-20-00708

Steinberg, A., Lyden, P. D., & Davis, A. P. (2022). Bias in stroke evaluation: Rethinking the cookie theft picture. *Stroke*, *53*(6), 2123–2125. https://doi.org/10.1161/STROKEAHA.121.038515

Stockbridge, M. D., Kelly, L., Newman-Norlund, S., White, B., Bourgeois, M., Rothermel, E., Fridriksson, J., Lyden, P. D., & Hillis, A. E. (2024). New picture stimuli for the NIH Stroke Scale: A validation study. *Stroke*, *55*(2), 443–451. https://doi.org/10.1161/STROKEAHA.123.044384

Taler, V., & Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, *30*(5), 501–556. https://doi.org/10.1080/13803390701550128

Verfaillie, S. C. J., Witteman, J., Slot, R. E. R., Pruis, I. J., Vermaat, L. E. W., Prins, N. D., Schiller, N. O., Van De Wiel, M., Scheltens, P., Van Berckel, B. N. M., Van Der Flier, W. M., & Sikkes, S. A. M. (2019). High amyloid burden is associated with fewer specific words during spontaneous speech in individuals with subjective cognitive decline. *Neuropsychologia*, *131*, 184–192. https://doi.org/10.1016/j.neuropsychologia.2019.05.006

Yazu, H., Kong, A. P.-H., Yoshihata, H., & Okubo, K. (2022). Adaptation and validation of the main concept analysis of spoken discourse by native Japanese adults. *Clinical Linguistics & Phonetics*, *36*(1), 17–33. https://doi.org/10.1080/02699206.2021.1915385

Yorkston, K. M., & Beukelman, D. R. (1980). An analysis of connected speech samples of aphasic and normal speakers. *The Journal of Speech and Hearing Disorders*, *45*(1), 27–36. https://doi.org/10.1044/jshd.4501.27