APHASIOLOGY
https://doi.org/10.1080/02687038.2025.2552801

Theoretical and methodological challenges to define a cross-linguistic discourse core outcome set of measures for aphasia

Halima Sahraoui [a], Silvia Martínez-Ferreiro [b] and Eva Soroli [c]

[a]Neuropsycholinguistics Research Unit, LNPL U.R. 4156, University of Toulouse, Toulouse, France; [b]Gerontology & Geriatrics Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), SERGAS, Universidade da Coruña, A Coruña, Spain; [c]Savoirs, Textes & Langages Laboratory, UMR 8163, University of Lille, Lille, France

ABSTRACT

Background: Production-based discourse assessment is an ecologically valid approach for investigating whether PWA have efficient verbal communication skills despite language impairments. A substantial body of research has focused on macro- and micro-linguistic discourse analyses in various languages, predominantly in English, and across different discourse domains, including productivity, informational content, and morphosyntactic elaboration. Besides, cross-linguistic studies have significantly advanced our understanding of aphasia by examining how language-specific structural properties influence the manifestation of linguistic deficits.

Aim: Despite the large number of discourse studies and a growing interest in building databases and automated tools, researchers and clinicians still face difficulties in analysing various aspects of discourse in a systematic and cross-linguistically valid way. To date, there is no unified cross-linguistic core outcome set of discourse measures. Advancing cross-linguistic discourse assessment in aphasia requires addressing theoretical, methodological, and practical challenges to identify and select reliable and valid discourse outcome measures for various languages, considering clinicians' needs, expectations, and consensus-based directions. This paper addresses the first considerations and challenges for incorporating multi-level, cross-linguistic, and typological perspectives into discourse assessment, with the aim of elaborating a cross-linguistic discourse core outcome set of measures (CLD-COS). Key features of commonly used discourse measures and their cross-linguistic operability are discussed, taking into account potential validity, standardization, and interpretability across languages.

Clinical implications: Clinicians have a limited number of effective production-based measures for discourse assessment applicable across languages. Clinicians and PWA, particularly those in multilingual and multicultural contexts, will benefit from a unified CLD-COS of measures supported by relevant constructs. In light of the considerations regarding discourse analysis across various

ARTICLE HISTORY

Received 6 May 2025
Accepted 21 August 2025

KEYWORDS

Core outcome set; discourse measures; production-based assessment; cross-linguistic validity; multilingualism

CONTACT Halima Sahraoui halima.sahraoui@univ-tlse2.fr Neuropsycholinguistics Research Unit, LNPL UR 4156, University of Toulouse, 05, Allées Antonio Machado, Toulouse 31058, France

languages, and the clinical needs reported in recent studies, this discussion represents an initial effort to pinpoint key issues involving multi-level and multi-component discourse analysis methods supported by evidence- and consensus-based practices.

Introduction

As posited by Dietz and Boyle (2018a), a discourse core outcome set of measures (D-COS) is required to facilitate a more profound comprehension of treatment effects on PWA's discourse and communicative abilities. Following the *Collaboration of Aphasia Trialists* consensus statement by Ali et al. (2022), cross-linguistic assessment and core outcomes in aphasia research are key themes in the field in aphasia research, with the aim of improving data comparability and advancing multi-site aphasia studies in both monolingual and multilingual contexts. The primary objective of this contribution is to propose a converging framework that incorporates cross-linguistic considerations to develop a cross-linguistic discourse core outcome set (CLD-COS) of measures while addressing significant theoretical and methodological challenges.

In this respect, current challenges in assessing discourse production across languages are highlighted when using a standardized core outcome set of discourse measures for clinical purposes (diagnosis and treatment outcome), thus facilitating the establishment of an appropriate CLD-COS. To move in this direction, it is essential to follow the recommendations outlined by the *COnsensus-based Standards for the selection of health Measurement INstruments* initiative[1] (COSMIN, Mokkink et al., 2010; Prinsen et al., 2016), which offers a structured process for identifying and selecting the most appropriate set of measures for outcome evaluation.

As a preliminary step, COSMIN advocates that researchers should clearly define the construct to be measured (Prinsen et al., 2016, pp. 3–4). This leads to a necessary cross-cutting examination of the complex linguistic and psycholinguistic processes involved in assessing preserved and impaired discourse production and their interactions with typological features across languages.

Despite advances in aphasia research, the assessment of discourse production in persons with aphasia (PWA) remains fragmented, especially from a cross-linguistic perspective. Cross-linguistic studies reveal striking variability related to language-specific patterns in grammatical operations such as tense, aspect, and case assignment, as well as disruptions in lexico-semantic and syntactic organization. However, clinical practice still lacks a unified, theoretically grounded framework for capturing these complexities, anchoring discourse constructs in cross-linguistic evidence, and paving the way for identifying robust measures for a CLD-COS.

Existing discourse analysis methods commonly used in research are also examined. These include the *Bilingual Aphasia Test* (BAT), the *Quantitative Production Analysis* protocol (QPA), and other emerging automated analysis tools. They offer valuable insights but raise concerns about cross-linguistic applicability and heterogeneous measures that incorporate macro- and micro-linguistic levels, including productivity, fluency, information content, and morphosyntax.

Furthermore, from a clinical perspective, we foreground the crucial need to establish a CLD-COS for improving assessment and intervention strategies. However, significant challenges remain in integrating multi-level, multi-domain validity and selecting reliable cross-linguistic outcome measures related to productivity, content information, and morphosyntactic complexity.

Finally, avenues for future research also include developing a CLD-COS of measures that considers both the interpretability of measures within a broad, usage-based framework (Boye & Harder, 2012) and aligning these discourse measures with communication Outcome Measurement Instruments (OMIs).

Cross-linguistic variability in PWA's performance

Growing interest in cross-linguistic research highlights the need to identify linguistic measures that are suitable candidates for implementation in a potential cross-linguistic COS. However, discrepancies in assessed patterns can be observed across languages. Such variability is particularly pronounced at the micro-linguistic level, as evidenced by studies on language-dependent performance patterns, which reveal disparities affecting morphosyntactic and lexical-semantic structures.

Language-specific error patterns

Although not yet fully representative, the growing body of cross-linguistic studies in the field of language disorders plays a crucial role in describing, explaining, and refining hypotheses about the nature of underlying deficits and their language-specific manifestations (Bates et al., 1991; Crago et al., 2008; O'Connor et al., 2005; Paradis, 1988). Models of language use that account for the specific properties of a given communication system greatly benefit from cross-linguistic data, especially when dealing with language use in aphasia. Paradis (2001) posits that the structural diversity of languages determines the linguistic form of surface production and, in particular, the patterns of phonological, lexico-semantic, and morphosyntactic errors in PWA discourse production.

Indeed, cross-linguistic investigations in aphasiology have been crucial in moving beyond the classical syndromic approach by reconsidering patterns of omission and/or substitution errors in close relation to the structural properties of a language. In a study comparing English, Italian, and German data from people with fluent and non-fluent aphasia, Bates et al. (1987) showed that speakers of richly inflected languages exhibit different patterns of morphological errors compared to English-speaking individuals who use a morphologically less marked system. Given these findings, the validity of usual symptom-based classification systems when applied to morphologically rich languages is called into question.

The *Cross Language Aphasia Study* project (CLAS; Menn & Obler, 1990a; 1990b was one of the first large-scale systematic studies of connected discourse, incorporating discourse tasks across 14 languages. For each language, two individuals with agrammatic aphasia and two non-brain damaged (NBD) control speakers told the story of their illness, a fairy tale (*Little Red Riding Hood*), a narrative based on the *Cookie Theft* picture (from the Boston Diagnostic Aphasia Examination, BDAE, Goodglass & Kaplan, 1972), and a story prompted by a set of four picture sequences (from the Wechsler-Bellevue stories: *farmer, theft, picnic,*

wake up, Wechsler, 1981). Analyses reveal that the omission of bound morphemes and the use of infinitive verb forms vary as a function of specificities of different grammatical systems and the morphosyntactic processing load associated with target inflected forms. While free-standing functional elements tend to be omitted across languages, the omission of bound grammatical morphemes is more variable. In highly inflected languages, such as Romance languages or Finnish, substitution errors were more frequent than omissions due to the complexity of inflectional paradigms. The simpler the paradigm, the fewer errors, as less morphosyntactic computation is required.[2] The study showed that canonical word order preferences differ as well: languages with fixed word order do not allow much flexibility, whereas languages with more flexible syntax (e.g., Finnish, Polish) allow for original yet grammatically correct word orders. PWA in Finnish and Polish often adopt structures with specific verb positions (initial in Finnish, final in Polish), resulting in idiosyncratic yet systematic patterns adapted to discourse requirements. Similarly, German-speaking PWA tend to avoid verb-final structures, which require more complex syntactic computations.

Besides, in their review dealing with various types of tasks (e.g., sentence production, comprehension, grammaticality judgment), Bates et al. (1991) argue that language-specific features predict cross-linguistic differences (within a given deficit), and within-language similarities (between people exhibiting different deficits). They emphasize that discourse variation across languages accounts for more variance than patient group distinctions. Moreover, individual performance also depends on preserved linguistic knowledge (e.g., morphosyntactic, lexical, and pragmatic), which directly influences the accuracy of surface-level output. Additionally, typological differences also lead to specific patterns of impairment in inflection or function word usage, yielding substantial variation from one language to another, and distinct patterns of fragile *versus* preserved morphology within one language, across language types, and populations. For instance, the accuracy of noun and verb production, the ability to exploit both grammatical and lexical redundancy, and the nature of substitution errors in discourse largely vary depending on language type.

Cross-linguistic research has reinforced the idea that language-specific morphosyntactic but also semantic features systematically shape the nature and distribution of error and compensatory strategy patterns in aphasia, highlighting the interplay between structural complexity and processing demands, as demonstrated in more recent research (Soroli et al., 2012; Thompson et al., 2013, among others).

Grammatical abilities: Tense/aspect and case assignment

As stated by Bastiaanse et al. (2011), tense and aspect are encoded through various language-specific strategies, making them ideal domains for studying time reference abilities and accessibility to their sub-components. For instance, English, Dutch, and Spanish combine simple verbs and periphrastic forms in this domain, whereas Russian, Greek, and Turkish rely primarily on simple verb forms. Besides, languages such as Chinese, Indonesian, and Thai use free-standing morphemes (like adverbs). In Bastiaanse et al. (2011) cross-linguistic study, Chinese, English, and Turkish agrammatic sentence production was scored according to whether the required verb inflection or free-standing morpheme was complete and correctly produced, with omissions or

substitutions of bound overt affixes or free morphemes (auxiliaries "be", "have", modal and aspectual verbs, and aspectual adverbs in Chinese) marked as errors, and lexico-semantic paraphasias were disregarded. The scoring system of the *Test for Assessing Reference of Time* (TART; Bastiaanse et al., 2008) was developed following these criteria to analyse a wide range of morphosyntactic time reference encoding means across 15 different languages, focusing on past, present and future reference in the active form in a close-to-discourse sentence production task. This method enabled fine-grained analyses of complex (language-specific) time reference expressions both in production and comprehension. Results and interpretation arising therefrom assume an underlying deficit affecting past reference when expressed through grammatical morphology, which manifests both in the production and comprehension of verb inflection and aspectual adverbs in agrammatic aphasia. This pattern reflects a selective impairment affecting past reference (captured by the *PAst DIscourse LInking Hypothesis*, PADILIH; Bastiaanse et al., 2011). Although experimental tasks were used, that is, verbs were not produced in connected and spontaneous spoken discourse, the TART procedure still involves processing of time reference, which has to be located in time frames as in connected discourse, and the scoring method proves to be topologically consistent and easily replicable across very different languages. Overall, results from other languages remain partially consistent with the PADILIH, as shown by the absence of this pattern for Italian PWA (Fyndanis et al., 2018). Moreover, varying results for future time reference are still an open issue (see e.g., Martínez-Ferreiro & Bastiaanse, 2013, for Catalan and Spanish).

Other cross-linguistic investigations compared different morphological patterns of case assignment and their realisation in Dutch and German agrammatic spontaneous speech and picture description. Ruigendijk et al. (1999) have shown that neither pronoun nor determiner case marking could occur without an overtly produced verb. This finding suggests that omission errors affecting pronouns or determiners in agrammatic aphasia are conditioned by the presence or absence of a verb in the syntactic structure, whether the case marking paradigm is simple or complex. This supports a syntactic, rather than a morphological, explanation for omission error patterns and points towards cascade effects. Depending on the case assignment system of a given language (grading from zero-marking to complex overt morphosyntactic marking), obligatory omitted pronouns and determiners may not be consistent enough to stand as relevant cross-linguistic core indicators of the type and severity of a person's aphasia. Nevertheless, there is still the option of drawing specific and within-language norms for typical case assignment in languages where pronouns and determiners are frequently omitted, and can stand, within a language, as core indicators of the type and severity of language disorders.

Lexico-semantic and syntactic abilities

In addition to morphosyntactic encoding and case marking, lexico-semantic and syntactic abilities exhibit considerable cross-linguistic variability that directly impacts performance in aphasia. According to Talmy's framework Talmy (2000), languages differ substantially in the options they offer for information encoding, semantic and syntactic packaging, For instance, in the domain of motion event encoding, *Manner* and *Path* components of motion are lexicalised differently depending on the typological profile of the language: Verb-framed languages such as French, typically encode *Path* within the main verb

leaving *Manner* in the periphery of the sentence or completely omitted, whereas Satellite-framed languages such as English, express *Manner* in the verb, systematically combined with *Path* adjuncts in compact and syntactically dense structures. Such typological differences in semantic component selection and syntactic packaging of information seem to influence not only how speakers linguistically describe events verbally but also influence the underlying cognitive processes involved in event construal, such as attention allocation and conceptual categorisation (Soroli, 2018, 2024). Notably, as demonstrated by Soroli (2011) and Soroli et al. (2012), experimental data in sentence production showed that PWA with similar clinical profiles (e.g., agrammatic aphasia) but from typologically different linguistic backgrounds (e.g., English- vs. French-speaking individuals) tend to adopt very different compensatory strategies that follow directly from the typological constraints of their mother tongues: English-speaking PWA focus more heavily on non-prototypical positions such as prepositional phrases, locatives or adverbial expressions, peripheral to the verb, to express the semantic component which is otherwise central and lexicalized in their language (e.g., "*Quickly quickly to the other side*". *Indeed, they* tend to use a *Manner* adverbial, instead of a *Manner* verb, combined with a *Path* prepositional phrase to express an event (e.g. a "run across the street" event), as opposed to French-speaking PWA who tend to opt for light verbs and the use of nominalizations to utter the core component in their language (*Path*), systematically omitting any *Manner* information (e.g., *"Il fait gauche droite", "He does left right"*: light verb combined with *Path* nominals to express the same "run across the street" event).

Additional evidence from cross-linguistic research reinforces this idea of the influence of typology on lexical selection in discourse. For example, Sung et al. (2016) compared the narrative discourse productions elicited from picture description in Korean and English, involving both neurotypical speakers and individuals with aphasia (anomic and Broca's). They found that Korean PWA, operating in a verb-salient language, produced more verbs and fewer noun phrases than their English-speaking counterparts. These findings show that the noun-verb dissociation classically used to distinguish aphasia subtypes must be cautiously nuanced cross-linguistically.

Finally, studies involving bilingual PWA highlight the added complexity of typological and experimental factors. Behavioural and eye-tracking paradigms reveal asymmetries in sentence processing and monitoring across language pairs. For example, Basque-Spanish bilingual PWA display differential comprehension performances depending on the structural features of each of their languages (Arantzeta et al., 2019). Similarly, Turkish-German bilingual PWA show better comprehension accuracy when processing subject and object wh-questions, with performance modulated by other factors, such as age of bilingualism onset or premorbid language use such as proficiency and exposure history (Arslan & Felser, 2018). These trends are further supported by meta-analytic evidence (Kuzmina et al., 2019), underscoring the necessity of considering bilingual and typological profiles in cross-linguistic discourse assessments.

Besides micro-linguistic features, including morphosyntactic and lexico-semantic phenomena, comprehensive discourse assessment also requires the integration of macro-linguistic indicators, including narrative coherence, informativeness, and structural organization. Studies in cross-linguistic language acquisition have long emphasized the importance of discourse-level structures, including narrative coherence, event sequencing, and information packaging, for understanding how speakers, children or adults, with

and without language impairments, encode and organize information across typologically diverse languages (Berman & Slobin, 1994; Hickmann, 2003; Strömqvist & Verhoeven, 2004). These findings underline the significance of both language-specific and universal constraints on discourse development. Building on this perspective, recent work has shown that macro-structural discourse features, such as narrative coherence, causal connectivity, and referential organization, are equally critical for characterizing discourse production in clinical populations, including individuals with aphasia (Hickmann & Soroli, 2015).

Integrating both micro- and macro-level assessments is essential to comprehensively capture the multidimensional nature of discourse in aphasia across languages. This dual-level approach enables more accurate profiling of discourse impairments and supports the development of cross-linguistically valid outcome measures.

Existing discourse scoring systems and measures used across languages

Since the 1980s, several key cross-linguistic discourse analysis protocols have been developed, yielding a set of quantitative measures most commonly used or adapted in a variety of linguistic contexts and languages. Among these, three protocols are particularly noteworthy, due to their extensive application across monolingual and multilingual aphasia corpora: the *Bilingual Aphasia Test* (BAT, Paradis & Libben, 1987), the *Quantitative Production Analysis* protocol (QPA, Saffran et al., 1989), and the EVAL program (Forbes et al., 2012) (see Appendix A). In these protocols, discourse is collected in various ways, being prompted generally *via* a semi-structured interview or storytelling based on strips, as in the BAT, yielding a minimum expected quantity of information produced. The QPA protocol involves a minimum of 150 core narrative words produced for a well-known story (*Cinderella*), excluding words outside of this core. Although sample length influences confidence in discourse analysis, recommendations change across protocols. The range of recommendations extends from a minimum number of produced words, with some studies proposing a minimum of 300 words for aphasia (Prins & Bastiaanse, 2004) and 700 words for dementia (Ossewaarde et al., 2020).

Discourse scoring and cross-linguistic equivalence in the bilingual aphasia test (BAT)

Unlike most language protocols originally designed for monolingual assessment, the cross-linguistic BAT discourse scoring system[3] (Paradis & Libben, 1987, p. 191, p. 214) has been conceived specially for bilingual speakers. It evaluates both spontaneous speech (e.g., illness history, occupation, living abroad) and picture-based narratives (e.g., the "nest story") in PWA. The BAT combines an initial subjective on-line rating, assessing quantity, fluency/speech rate, pronunciation, grammar and vocabulary on a 1 "poor" to 4 "normal" scale), combined with a more detailed offline scoring system comprising 22 linguistic production-based measures and three other rating scales (Appendix A provides an overview of these 22 variables alongside a discourse variable taxonomy provided by Bryant et al., 2016).

The BAT's flexibility has allowed its application beyond its original scope. The spontaneous speech scoring system now extends to different types of connected spoken discourse tasks, as in Nilipour's (2000) study of Persian agrammatic aphasia, which

includes personal narratives, the *Cookie Theft* picture description task, and the Wechsler-Bellevue cartoon, as adapted for the CLAS study. It has also been applied to other neurogenic conditions: Zanini et al. (2010), for example, used a subset of the BAT scoring for the analysis of spontaneous speech in people with Parkinson's Disease, revealing language-specific error patterns with higher morphological vulnerability in L1-Fruilian than in L2-Italian.

As emphasized by Paradis (2011, p. 430), any adaptation or use of the BAT in different languages has to satisfy the principle of "cross-linguistic equivalence", depending on the quality of cross-systems adaptation (e.g., selecting a relevant set of minimal pairs as stimuli for each language system, rather than simply translating words from one language to the other). The cross-linguistic equivalence principle, which should also be applied to discourse scoring and measures, is addressed only for the type-token ratio (TTR) which has to be "equivalent across languages irrespective of language type" (e.g., agglutinative languages, languages with no free-standing determiners or copulas), by counting types according to the corresponding entry in a referenced monolingual dictionary (minimum 40.000 entries) (Paradis & Libben, 1987, pp. 28–30, p. 121). Additionally, to deal with morphosyntactic variability across languages, closed-class words (determiners, particles, prepositions, conjunctions, copula, pronouns) are excluded from the type counting and the TTR calculation. Missing pre- or post-posed obligatory morphosyntactic morphemes, either unbound function words or inflectional affixes (person, tense, aspect, gender, plural, etc.), are not distinguished. Reducing the quantification to errors in a single broad grammatical-morphemic category is a simple and effective solution to preserve cross-linguistic equivalence, though morphosyntactic complexity is thus only partially captured.

Currently, the BAT discourse scoring system provides standardized procedures only for mean length of utterances (MLU), TTR, and number of verbs per utterance, making it a relatively comprehensive, highly flexible, and adaptable tool across languages. However, a notable limitation is the absence of published normative and cross-linguistic valid discourse data, which hinders its clinical generalizability.

The quantitative production analysis protocol

The *Quantitative Production Analysis* protocol (QPA, Berndt et al., 2000; Rochon et al., 2000; Saffran et al., 1989) was first designed to characterise formal linguistic aspects of agrammatic aphasia in narrative discourse. The QPA procedure is clinically sensitive as it helps to distinguish between fluent and non-fluent aphasia (Bird & Franklin, 1996). The scoring system focuses on lexical and morphosyntactic phenomena at the word and utterance levels, differentiating various profiles of aphasia subtypes and degrees rather than agrammatic aphasia alone (Gordon, 2006). More recently, this procedure has also been applied to other syndromes such as variants of primary progressive aphasia (Lavoie et al., 2021, 2022). In the original QPA protocol, transcription and annotation are processed manually and based on framed instructions, yielding a broad range of scores and possible calculation extensions. The annotation procedure can be easily adapted from English to typologically closely related languages, as demonstrated by the complete adaptation to French (Sahraoui, 2009; Sahraoui & Nespoulous, 2012). Yet, its cross-linguistic validity for more distant language families remains to be empirically confirmed, and the appropriateness of

its measures (see Appendix B) still requires validation across typologically distant languages.

Automated measures for discourse evaluation

AphasiaBank (MacWhinney et al., 2010, 2011) is an international database of discourse and interactions for the study of aphasia[4] which provides a methodological framework for clinical language research, facilitating corpora sharing, interoperability, and data accessibility. The cross-linguistic corpora of AphasiaBank are elicited using a standardised protocol adapted to different languages. Developed by MacWhinney (2000), the transcription and coding system (human speaker coding in CHAT format) combined with a specialized *Computerized Language Analysis* (CLAN) program, offer the possibility for a wide range of quantitative analyses based on standardised coding principles and measures (e.g., frequency calculations, TTR, MLU). These semi-automated analyses can be applied, for instance, to the study of repetitions, false starts, number of verbs, nouns, functional words, collocations, or basic fluency indicators, with the potential to create new coding scripts for additional dependent variable analysis. Since approximately 2010, the growing number of studies published in the field of discourse in aphasia and the prominence of corpus linguistics methods have been relying on shared and accessible databases, together with standardised and semi-automated coding and analysis tools adaptable to languages other than English.

Based on the CHAT-CLAN program, the EVAL system has been developed for clinicians (Forbes et al., 2012, 2014), enabling quantified comparisons across individuals or groups selected from the database across 34 discourse production measures (see EVAL measures in Appendix A). Counting and calculations are essentially related to utterance and word/lemma counting, morphosyntactic tagging, chronometric measures of utterances, fluency, and error phenomena. The EVAL measures are fully sensitive and relevant in English, but some of them are not directly transposable to other languages, especially when drawing measures from cross-linguistic data, either in research or clinical settings. One limitation of the EVAL program is the distinction between open- and closed-class words and related measurements, a classical morphemic distinction which is problematic as this dichotomy is not retained as such in recent theories of functional grammar and usage-based frameworks applied to discourse analysis in aphasia (Boye & Harder, 2012; Boye et al., 2015). However, the automation of linguistic information coding and extraction using CLAN can be extended with supplementary coding, making the system highly flexible and adaptable to several languages and theory-driven annotations within the CHAT formats.

Moreover, the CLAN program offers a set of automated measures implemented from previous reliable and extensively used methods such as the QPA (Saffran et al., 1989; see in; Fromm et al., 2021), the *Main Concepts Analysis* or *Correct Information Unit* analysis (Capilouto et al., 2005; Nicholas & Brookshire, 1993, 1995), the *Northwestern Narrative Language Analysis* (originally developed by Fromm, Macwhinney, et al., 2020) and the *Core Lexicon* measures (S. G. Dalton & Richardson, 2015; S. G. H. Dalton et al., 2020; Kim & Wright, 2020), that go beyond micro-analysis and extend to macro-level considerations. In this direction, extensive coding can be added to standardised transcripts, including multi-level coding and scoring of discourse units for targeted discourse content and cohesion

(such as the *Story Grammar* or the *Main concept – Sequencing & Story Grammar*, MSSG,[5] Richardson et al., 2021). The outcome measures of information content highly depend on human coding, a crucial phase of data pre-processing consisting of annotating the quality of the information produced in regard to the expected narrative scheme and information content.

CLAN methods for transcribing, scoring, and analysing different discourse levels in aphasia, including linguistic and temporal aspects of (non)fluency, have been automated based on English-language data. Nevertheless, efforts are underway to transfer and use automated tools across different languages. Notably, this requires particular attention to the development of multilingual natural language processing algorithms operable within CLAN to improve the efficiency of automated transcription and coding (H. Liu et al., 2023), and to develop new automated fluency assessment methods dedicated to aphasia in various languages (Fontan et al., 2023; J. Liu et al., 2023).

Key clinical need for a discourse and functional COS of measures

Assessment batteries, as well as clinical linguistics more broadly, place particular attention on discourse (e.g., spontaneous or semi-spontaneous speech, narrative, picture description), as it represents "the most meaningful, natural, ecologically valid, and available variety of communication" (Fromm et al., 2020, p. 2). Indeed, discourse is a functional manifestation of language use close to real-life communicative contexts, offering insight into language abilities beyond isolated linguistic units. Traditional comprehensive assessment batteries and their potential cross-linguistic adaptations typically incorporate one or more spoken discourse tasks, e.g., the *Boston Diagnostic Aphasia Examination* (BDAE; Goodglass & Kaplan, 1972), the *Bilingual Aphasia Test* (BAT; Paradis & Libben, 1987), the *Comprehensive Aphasia Test* (CAT; Swinburn et al., 2012), the *Aachen Aphasia Test* (AAT; Huber et al., 1983), the *Montréal-Toulouse Protocol* (MT-86; Lecours et al., 1996), or the *Grémots* (Bezy et al., 2016), among others, which are widely used in clinical practice.

Besides, protocols dedicated to aphasia discourse analysis in languages other than English are gaining visibility. Such protocols include, among others, the *Análisis del Lenguaje Espontáneo en Adultos* protocol (ALEA; Méndez-Orellana et al., 2022); the *Analyse voor SpontaneTaal bij Afasie*, originally designed for Dutch (ASTA; Boxum et al., 2013; Van der Scheer et al., 2011), and the *Amsterdam-Nijmegen Everyday Language Test* (ANELT; Ruiter et al., 2011, 2022) for Dutch as well. These protocols prioritize clinically feasible yet linguistically sensitive analysis. In particular, the ALEA was developed upon request for speech and language therapists (SLT) working with Spanish-speaking PWA, and based on the QPA protocol (Saffran et al., 1989) originally designed for English and non-fluent aphasia. It comprises various linguistic measures such as MLU, paraphasia frequency, and measures of grammatical accuracy and complexity, and allows for a simple codification of categories of interest, such as nouns and verbs, carefully described in the transcription and coding guidelines. The ALEA was designed to balance data collection time and depth of analysis according to the specificities of Spanish and the clinical and cultural contexts in which it is going to be used, ensuring applicability in daily practice.

Therefore, discourse analysis serves as a critical diagnostic window, enabling assessment of language performance and rehabilitation progress related to communication abilities. Fundamental and clinical research in aphasiology has long and continuously

adopted a pragmatic perspective, with the communicative intent and understanding of meaning being of primary interest. In line with this perspective, the main goal in therapy is to optimize PWA communication skills by implementing discourse-based and conversational interventions that target functional outcomes (A. Holland et al., 2019; A. L. Holland, 2021; see also Goldberg et al., 2012, dealing with conversational script training intervention). Building upon Audrey Holland's influential work to advance discourse- and communication-based therapy in aphasia, Armstrong and Hersh (2024), p. 360) advocate for a functional approach to aphasiology, in which "linguistic discourse analysis enables the unpacking of functional communication and gives us a framework for looking at it systematically". This call is supported by previous findings from a major clinical practice survey conducted by Stark et al. (2021) in English-speaking countries (USA, UK, Australia), which revealed that SLT often lack systematic guidelines for analysing discourse data at structural or functional levels. The authors emphasize the need for standardized annotation procedures and core outcome measures with high-quality psychometric properties, supported by normative data for efficient assessment and treatment. Stark and colleagues recommend the development of standardised pre- and post-treatment, multi-level discourse measures tailored to therapy interventions that span multiple discourse levels (word, sentence, and macro-structure) and various types of spoken discourse tasks. Indeed, as shown by Whitworth et al. (2015), a better knowledge of discourse-genre specificities helps enhance clinical interventions in everyday speaking contexts, including recount, procedural, expository, and narrative discourse. The study revealed that healthy adults consistently used routines that support coherence and cohesion in speech, aiding both comprehension and production. This provides a tangible framework for clinicians to assess impairments and guide interventions to improve everyday communication.

To achieve these goals, several teams, such as the FOQUS Aphasia working group of the *Collaboration of Aphasia Trialists* network, point out the necessity to exploit large databases to establish standard guidelines and enhance the consistency of discourse measures reported in aphasia research. Moreover, findings from Bryant et al. (2017) demonstrate that clinicians view discourse analysis as highly relevant to intervention, yet practical limitations, such as time constraints and limited training, often result in reliance on clinical judgement methods over fine-grained transcription-based analyses. These findings lead the authors to call for research efforts aimed at adapting linguistic discourse analyses into feasible clinical applications, through the development of more accessible and effective core outcome sets of discourse measures, and enhanced training for clinicians (Kintz & Wright, 2018).

In a similar vein, Dietz and Boyle (2018b) suggest that effective discourse core outcome sets (D-COS) must integrate both micro-structure and macro-structure levels of scoring and analyses. They underscore the need for consensus around D-COS frameworks for both research and clinical contexts. Responding to this imperative, various studies have highlighted key clinical requirements: establishing standardised scoring and analytic methods for discourse at different structural levels, ensuring psychometric validation, and defining guidelines for selecting appropriate discourse measures from a common set of tools. Furthermore, a clear demand emerges for the development of novel clinical instruments capable of evaluating discourse within a cross-linguistic and multilingual context.

Key challenges for integrating multi-level and multi-domain validity of cross-linguistic discourse measures

Multi-level discourse construct

Linguistics traditionally distinguishes between different levels of discourse organization, from surface structures to underlying forms, meanings, and actions, encompassing both verbal and non-verbal dimensions (Van Dijk, 1997). According to Van Dijk (1997, p. 5), discourse must be understood as language use in a communicative context, involving interrelated "components" which are "ordered" in a specific way, and "combined in larger constructs". Discourse thus comprises multiple layers focusing on a set of "various structures" sequenced in a discursively meaningful and coherent larger construct. Indeed, discourse relies on interrelated *components* (phonemes, syllables, word, phrases, propositional or clausal unit, sentence or sentential unit, prosodic unit, text unit, etc.) which are organized so that natural language listeners and speakers understand and express contents, and interpret links between contents within a specific context (e.g., using time or person reference marks, discourse particles, textual schemes).

Besides, discourse meaning relies on both *micro-level analysis* (order of words and sentences or propositions, instantiating local coherence) and *macro-level analysis* (global coherence through sentence sequencing and ordering, speech acts and interaction structures. Van Dijk (1997, p. 30) further distinguishes two analytical perspectives: that of the analyst (e.g., the researcher in clinical linguistics or the speech and language therapist), who decomposes discourse into structured components, and that of the natural language user (the listener/speaker, including the person with language disorders), who integrates all levels and components into a strategic activity to communicate efficiently.

As suggested by Armstrong (2000), researchers in aphasiology should aim to connect micro- and macro-linguistic aspects to achieve a holistic, comprehensive understanding of aphasic discourse, considering methodological consistency, particularly in language sampling, and the different discourse genres to enhance the validity of future studies.

Previous work has already shown that a multi-level qualitative and quantitative discourse analysis approach is essential for assessing discourse abilities in aphasia and valuable for developing targeted interventions. For instance, Marini et al. (2011) demonstrated the effectiveness of this approach by assessing productivity, errors, narrative organisation/local and global coherence, and informativeness measures in two case studies in Italian. Similarly, Wright and Capilouto (2012) highlighted the importance of combining local and global coherence measures in their investigations.

Multi-domain discourse construct

Such research directions have generated a large body of work aimed at identifying relevant and robust discourse variables in different types of discourse (Bryant et al., 2016; Pritchard et al., 2017, 2018). In particular, Bryant et al. (2016) conducted a comprehensive review of 165 studies spanning four decades that employed linguistic discourse analysis to assess language performance in PWA across different discourse tasks (expository, descriptive, and narrative discourse, the latter type being the most commonly used task, featuring in 101 studies).

Their review identified a total of 536 different discourse measures (both raw and ratios), which can be grouped into three broad categories of variables: verbal productivity, information content, and grammatical complexity. These domains, assessed through both qualitative scoring and quantitative analysis, were found to be instrumental in distinguishing between fluent and non-fluent aphasia. For example, productivity measures, such as speech rate, frequency of pauses, dysfluencies, and total word output, are informative for evaluating fluency. Besides, severity is also largely determined by the type and frequency of errors, including those related to information content, lexico-semantic accuracy, discourse cohesion, and morphosyntactic aspects.

These different domains of measures also provide valuable insights to characterize potential compensatory strategies that may be employed by PWA at various linguistic levels. Such analysis helps to contextualize discourse abilities within the individual's language profile, task-specific demands, and typological characteristics of the language(s) spoken (see Appendix A).

Among the measures and analyses protocols reviewed, the *Quantitative Production Analysis* procedure (QPA, Saffran et al., 1989) emerged as the most frequently used, appearing in 17 out of 165 studies. Even though Bryant et al. (2016) review focused exclusively on studies involving English-speaking participants, the findings remain substantial for the development of a cross-linguistic discourse core outcome set of measures (CLD-COS). In light of current evidence, it appears essential that any discourse COS should incorporate, at a minimum, relevant measures of productivity, information content, and grammatical complexity. These dimensions represent key dimensions of discourse organisation and are critical for capturing both inter-individual and task-related variability from multi-domain perspectives at both micro- and macro-linguistic levels.

Variability and multi-dimensional discourse construct

Discourse variability also requires particular attention. As Armstrong (2018) noted, quantifying connected speech and discourse is a challenging endeavour because the objective is to identify discourse measures that demonstrate high stability,[6] reliability, and validity,[7] despite variability being a fundamental aspect of both typical and atypical language use. A vast array of studies has already shown that variability is the hallmark of language behaviour across different discourse genres and levels of performativity, as well as across individuals and communication settings (Bastiaanse, 1995; Nespoulous, 2000; Kolk, 2007, among others). As demonstrated by Sahraoui and Nespoulous (2012)'s study focusing on agrammatic discourse production, PWA may exhibit different patterns of impaired and preserved abilities for the same linguistic variable, depending notably on how discourse is adapted and managed under communication settings or discourse types. The variability may be influenced by task instructions, as supported by the adaptation theory framework (Hofstede & Kolk, 1994; Kolk, 2006), the anticipated discourse format (e.g., monologue *versus* conversation settings, structured vs. unstructured spoken discourse, see Leaman & Edmonds, 2021a, 2023), or the type of elicitation task employed (e.g., picture description *versus* storytelling, which can yield varying morphosyntactic accuracy depending on determiner or auxiliary cues, see Schnur & Wang's, 2024).

A key issue in discourse analysis for diagnosis and therapy is that it requires the use of qualitative discourse-related metrics, based on theory-driven constructs derived from

linguistic and psycholinguistic concepts and categories. That is to say, language (and processing) is a naturalistic and complex manifestation of communicative behaviour which needs to be transformed into clinically relevant, measurable units. One major difficulty is that discourse assessment protocols must ensure that these metrics are, first, theoretically relevant, and second, sensitive, reliable, and clinically effective.

The validity of any discourse-related measure is deeply anchored in corpus linguistics methods, discourse analysis theory, and psycholinguistic models. The validity standards of discourse measures, therefore, depend on their multi-level, multi-component, and cross-linguistic adequacy.

Thus, content, construct, and structural validity become particularly challenging when dealing with discourse production, as it reflects different and interrelated processes subject to variability. Indeed, construct validity for optimal cross-linguistic discourse measures requires a threefold adequacy: first, structural validity across linguistic levels (micro or macro); second, the dimension reflected (productivity, information content, or morphosyntactic complexity); and third, the extent to which the measure achieves cross-linguistic equivalence.

Therefore, a key challenge is to develop a comprehensive discourse COS that captures the most clinically relevant components for aphasia assessment, encompassing both micro- and macro-linguistic levels of discourse, while accounting for the integrative processing demands of managing language impairments concerning individual, task, and typological variability.

Ensuring the interpretability and clinical applicability of discourse measures across levels of analysis, by COSMIN standards, is thus essential.[8] For a discourse COS of measures to attain cross-linguistic validity, the micro-linguistic level, comprising phonological, morphological, semantic, and syntactic structures, must be prioritized and reinforced. This is particularly important because these elements are more sensitive to typological variation than macro-level information content and contextual meaning structures.

Consequently, the integration of a multi-dimensional assessment method means considering the complex nature of discursive abilities in aphasia. To ensure that the assessment method remains comprehensive yet feasible, crucial points need to be discussed in depth and further: Which set of discourse measures best reflects the multi-dimensional nature of impaired and preserved abilities in PWA's discourse production, and how can they differentiate affected from preserved linguistic levels or components? In terms of redundancy, which measures should be excluded from a discourse COS of measures because they reflect overlapping constructs?

Challenges for identifying and selecting reliable and valid cross-linguistic discourse outcome measures

Cross-linguistic relevance of frequently used measures in aphasia research

Historically, selecting relevant measures showing cross-linguistic validity has not necessarily been at the forefront of aphasia clinical research. It is now necessary to address the question of the validity of discourse measures from a cross-linguistic perspective to guide the choice of theoretically sound metrics. The COSMIN standards defining the taxonomy and measurement properties of health-related patient-reported outcomes (Mokkink et al.,

2010) include cross-cultural validity, when dealing, for example, with cross-cultural adaptation of test stimuli or questionnaires. While COSMIN emphasizes adaptation and translation, it is important to specify that our focus here is more specifically on cross-linguistic validation of the discourse measures themselves.

For example, the BAT has proposed the "principle of equivalence" when applying cross-linguistic discourse measures for assessment (upon the multilingual intuition of the analyst). However, the construct of cross-linguistic equivalence has not yet been subject to any cross-linguistic psychometric validation using production-based analyses, whether in the BAT or any other comparative framework. We propose that the cross-linguistic validity of a discourse measure should be evidenced through production-based analysis. Thus, calid construct selection must align with both general psychometric standards and cross-linguistic equivalence, and special attention must be given to morphosyntax at the micro-linguistic level, where cross-linguistic variability presents the greatest challenge.

Other recent contributions illustrate the challenges in identifying and selecting psychometrically sound discourse outcome measures in other less-documented languages, such as Alyahya's (2024) study for Arabic (including three discourse tasks: picture description, picture storytelling and procedural); or Boucher et al. (2022) study for French-Canadian including the picnic scene from the Western Aphasia Battery-Revised (Kertesz, 2006), see in Appendix B the set of measures retained in both studies).

Productivity and fluency measures

Fluency remains a contentious construct in aphasia research. Gordon (1998) highlighted the difficulties in reaching agreement between clinicians on the concept of fluency and the associated indicators for its clinical assessment based on different BDAE subtests, including the discourse task (*Cookie Theft*). Such inconsistencies stem from the multidimensional nature of the concept of fluency, variably assessed by clinicians through phonemic, lexical, and morphosyntactic accuracy, or speech timing and effort (speech rate, dysfluencies shown by pausing, hesitations, and self-corrections).

In a recent study, Gordon and Clough (2024) showed that clinicians' perception of fluency correlated most strongly with objective measures of speech rate and utterance length. Interestingly, these findings are consistent with a scoping review (2012–2022) by Cordella et al. (2024) that examined quantitative methods used to characterise connected speech fluency in aphasia research focusing on post-stroke aphasia (PSA) and primary progressive aphasia (PPA). Despite a wide variety of different fluency measures reported in the 45 included studies (85% of which were English-based), resulting in 209 different quantitative speech and language features, the results show a consensus on the most commonly used key measures: speech rate and total word count across aetiologies, together with MLU in PSA studies.

These findings provide preliminary guidance for clinicians and researchers seeking to incorporate quantitative indicators of fluency into the assessment process, facilitating more accurate and standardised assessments. Although focused on English, productivity measures like speech rate and MLU, which is a measure at the interface of productivity and morphosyntactic complexity, remain promising candidates for a cross-linguistic discourse COS and should therefore be retained. Besides, lexical diversity indices such

as TTR or textual lexical diversity (TLD) are also widely used in discourse analysis (MacCarthy, 2005; McCarthy & Jarvis, 2010) and validated for narrative aphasia assessments (Cunningham & Haley, 2020; Fergadiotis et al., 2013).

Content information and lexical measures

Pritchard et al. (2017) scoping review showed that among discourse information measures, and despite heterogeneity from discourse data types, scoring methods, and statistical analysis, main concepts and single word information measures scored as Content Information Units, CIUs) are among the most valid and reliable. CIUs scored at the macro-linguistic level (isolated and connected propositions) or micro-linguistic level (lexical content), should thus be prioritized as potential candidates for a CLD-COS, in line with widely used methods outlined in a large number of studies focusing on this domain of measures (Capilouto et al. 2005; Fergadiotis & Wright 2011; Gordon 2008; Hameister & Nickels 2018; Kurland et al. 2023; Leaman & Edmonds 2021a; Leaman & Edmonds, 2021b; Nicholas & Brookshire 1993; Nicholas & Brookshire, 1995; Richardson & Dalton 2016, Richardson & Dalton 2020; Richardson et al. 2021; S. G. Dalton & Richardson 2015; Wright et al. 2010; Wright & Capilouto 2012).

Morphosyntactic complexity measures challenging cross-linguistic assessment

As was discussed previously in this article, cross-linguistic variability complicates the use of universal morphosyntactic measures, in particular at the micro-linguistic level. Overall interpretation of cross-linguistic data is challenging because morphosyntactic categories and processes are sensitive to typological variability. To date, there has been no consensus on the most effective cross-linguistic indicators. Yet, measures such as morphosyntactic accuracy (grammaticality), clause complexity (subclausal, clausal, multiclausal), and open- *versus* closed-class word ratios remain frequent in most studies. Moreover, fine-grained morphosyntactic complexity measures are particularly developed in the QPA protocol to capture in-depth morphosyntactic reductions (such as verb inflection index, auxiliary score, determiner or pronoun ratios, sentence elaboration index, Saffran et al., 1989; see Appendix A). Thus, a paramount challenge is to identify which micro- and macro-morphosyntactic measures are relevant across the board, and to define how accurately they reflect differential performance beyond typological variations of structurally diverse languages.

What do we expect from cross-linguistic discourse outcome measures?

Reliability standards

A substantial number of discourse measures have demonstrated good psychometric quality in numerous studies, particularly in English. Inter-rater and intra-rater coding reliability of the *Cookie Theft* discourse procedure was investigated by Powell (2006) using the protocol-related decision tree to guide and standardise utterance segmentation and coding, and utterance complexity (empty, subclausal, clausal, multiclausal). Powell found that overall scoring accuracy was low and recommended that coding systems

should systematically report scoring reliability by comparing examiners with different clinical experience levels, and include procedures provided to assessors with varying levels of expertise in speech and language therapy. These findings reflect the need to train SLT using standardised scoring systems, as also confirmed by Cruice et al. (2020) survey of discourse assessment and intervention in the UK. They also showed that while SLT are highly concerned with speech analysis, factors such as limited time, expertise, resources, and training hinder the effective use of speech analysis tools. Moreover, this study calls for the development and validation of a standardised discourse scoring and analysis protocol for implementation in routine clinical practice. This is of particular significance for all measures, including the calculation of the MLU.

Following the methods from Marini et al. (2011), Nicholas and Brookshire (1993), and Saffran et al. (1989), Alyahya (2024), focusing on Arabic language) revealed variability in the construct validity and reliability depending on discourse stimuli and the tested group. In particular, inter-rater reliability was not fully satisfactory for all discourse stimuli (TTR and proportion of nouns), and it varied across groups (for TTR and MLU). Notably, MLU demonstrated poor inter-rater reliability across all discourse stimuli within the neurotypical control group, but exhibited greater consistency within the aphasia groups. While MLU is widely used in research and clinical assessment as a reliable indicator of syntactic complexity in the connected speech of NBD speakers or PWA, Alyahya (2024) excluded it in favour of complete sentence measures because it was not sufficiently reliable. By contrast, Boucher et al.'s (2022) study focusing on French and using CLAN computations and coding instructions from Colin et al. (2016) demonstrated high inter-rater reliability for transcription and scoring, including for MLU.

These inconsistencies between studies from different languages highlight divergent approaches to assessing reliability. Rejecting measures like MLU based on inter-rater variability across tasks, as in Alyahya (2024) study, is questionable, as it neglects that such variability often reflects adaptive strategies, particularly in non-fluent aphasia, where morphosyntactic ellipsis is reflected by variability of MLU across tasks (Kolk, 2001, 2006). Measures influenced by discourse task type may capture important variations in productivity and complexity, as seen in differences between semi-spontaneous, narrative, and descriptive tasks with regard to MLU and other morphosyntactic measures (Sahraoui & Nespoulous, 2012).

Another key issue is the formalization of annotation procedures. Transcription, segmentation, and token-counting instructions are often insufficiently detailed, compromising reliability. Therefore, Stark et al. (2021) recommend comprehensive reporting of discourse scoring and segmentation methods. As suggested in Alyahya's (2024), discrepancies in utterance segmentation may have undermined some measures, even though protocols like the QPA have demonstrated strong reliability (Rochon et al., 2000; Saffran et al., 1989).

Selected discourse measures from different languages are not consistent, as shown when comparing which discourse measures have been selected in protocols dedicated to Arabic (Alyahya, 2024) and French (Boucher et al., 2022). Notably, in both sets of discourse measures reported (see Appendix B), essential variables such as verb tense and inflectional morphology are not addressed at all. Despite previous evidence indicating that verb inflection is a crucial feature in assessing PWAs' abilities across languages, this domain of discourse construct is missing in the reference data of both languages.

Treatment and goals-related reliability for evidence-based practice

To improve Evidence-Based Practice (EBP) by selecting relevant discourse outcome measures for intervention, Boyle (2020) recommends focusing on the relation of discourse measures to client intentions, the therapy and work settings, therapy goals (quality of life), and their psychometric properties. A set of questions is therefore proposed to clinicians to help them decide which discourse outcome measure should be relevant as a baseline variable, ensuring that fluctuations in scores effectively reflect micro- and macro-structural changes induced by treatment. By answering these targeted questions, clinicians have to match a given outcome measure to what is expected in the treatment, and specifically identify which aspect of discourse is expected to be improved (micro-structure and/or macro-structure, discourse genre), and its related psychometric properties, such as intra- and inter-rater reliability of at least 0.70[9]; test-retest reliability of at least 0.90; report of the standard error of measurement (SEM); report of minimal detectable change value (MDC) and effect size.

A key EBP systematic review was conducted by Dipper, Marshall, Boyle, Botting, et al. (2021), compiling an inventory of discourse treatment design quality, including the intervention reports, the range, type, and content of outcome measures used, and the treatment efficacy from evidence-based practice. Considering 514 different outcome measures reported across 25 studies, they found that words-in-discourse measures such as CIUs, content words, nouns, verbs, and adjectives are the most commonly used when evaluating discourse treatment efficacy. Moreover, two main conclusions regarding discourse measurement emerged from their review. First, they noted that words-in-discourse measures are numerous and highly heterogeneous. Second, they emphasised the need to identify further outcome measures at both the micro-structure and macro-structure levels. They concluded that word production extracted from discourse should be considered a salient core outcome measure. Finally, although the survey did not specifically address the cross-linguistic applicability of scoring methods, the proposed general guidelines can be used as a baseline, susceptible to being adapted, to improve clinical assessment and meet clinical needs effectively in a cross-linguistic perspective.

Besides, assessing discourse production in multilingual speakers in their different languages presents a significant challenge, particularly when considering the complex performance patterns of multilingual PWA and the heterogeneous outcome measures reported across studies (see Goral et al., 2023 for a review). Moreover, both single and multiple case studies illustrate the use of discourse-level and typologically motivated compensatory strategies in multilingual PWA (Penn & Beecham, 1992; Penn et al., 2001). Potential cross-linguistic generalization associated with therapeutic interventions is therefore evidenced through pre- and post-treatment assessments, highlighting the need for CLD-COS, which should provide standardized metrics for multilingual discourse assessment to inform both research and clinical practice.

Concerning CIU metrics, Conner et al. (2018) examined discourse across eight languages in multilingual PWA. Their findings indicate that language proficiency, beyond typological differences, significantly modulates the efficiency of language production in treated *versus* untreated languages, with generalization to non-treated languages. Interestingly, greater generalization was observed in higher-proficiency languages and, counterintuitively, in typologically more dissimilar languages. Moreover, lower proficiency

languages were associated with more frequent code-mixing. Nevertheless, the results are nuanced given the absence of consensus regarding optimal outcome measures for evaluating treatment efficacy across languages, which complicates interpretation.

Current cross-linguistic approaches for EBP often rely on a limited scope of analysis, such as CIU counts or code-switching frequency, with notable concerns regarding inter-rater reliability and the reduction of discourse analysis to a single component. These methodological constraints necessitate more refined and multilingual-compatible measures for more nuanced and comprehensive assessments of discourse phenomena reflecting recovery in multilingual PWA.

Implications and future directions

Discourse assessment across languages must be theoretically grounded and clinically applicable. Ideally, it should involve a limited yet comprehensive set of measures covering various discourse levels. This position paper advocates for the development of a cross-linguistic discourse core outcome set (CLD-COS), composed of production-based measures ensuring reliability and validity across different languages while accounting for both qualitative and quantitative variability.

Such a CLD-COS aims to provide multi-level discourse assessment, incorporating indices of verbal productivity, information content, and grammatical complexity. To support this, future research should review normative data across languages and initiate new corpus-based studies that meet standards of validity, reliability, and clinical sensitivity. Development of the CLD-COS also requires addressing cross-linguistic variability and ensuring that standardised measures are both cross-linguistically equivalent and clinically applicable to different targeted therapeutic goals. Such a framework should allow flexible discourse scoring and be adaptable within or across languages. A robust CLD-COS would therefore enhance the diagnosis, monitoring, and treatment of aphasia, improving outcomes for monolingual and multilingual individuals alike.

The interpretability of a CLD-COS of measures: usage-based and psycholinguistic adequacies

As highlighted by COSMIN, interpretability is critical for any measurement instrument. A key requirement of a valid CLD-COS is the interpretability of cross-linguistic descriptions concerning hypothesis validation about impaired and strategic discourse use, especially those grounded in recent research developments in functional discourse grammar theories applied to discourse produced by PWA. In this vein, it is first useful to rely on the "circularity effect" that intrinsically links theoretical assumptions and linguistic annotations (Consten & Loll, 2012). In this respect, theoretical assumptions and psycholinguistic interpretations are derived from annotated linguistic phenomena (because they are meant to reflect psycholinguistic discourse processing), reinforcing the need for valid and interpretable measures for PWA discourse-based assessment. Consten and Loll (2012, p. 711) also assume that "general functional categories are less canonized than structural ones". This assumption aligns with the fundamental principles of functional linguistic theories (Dik, 1997; Halliday, 1985). They postulate meaning, pragmatic, and cognitive adequacies in support of linguistic phenomena description and their variability when

applied in particular to the study of discourse with aphasia (Sahraoui, 2015). In line with the functional theory, recent usage-based approaches challenge conventional morpho-syntactic analyses and underscore the impact of typological specificities by reconsidering pre-conceived structures that have been canonized for analysis. In this way, discourse analysis transcends traditional grammatical categories (e.g., open *versus* closed class words). In light of this, the usage-based approach challenges the classical dichotomy between open and closed class words by analysing morphosyntactic components according to their discursive (primary *vs*. secondary) and their related cognitive status (Boye & Harder, 2012; Martínez-Ferreiro et al., 2020; and for an extensive review focusing on agrammatism, see; Boye et al., 2023). In support of this general theory of typical and atypical language use, cross-linguistic data from Danish and French (Boye et al., 2015; Nielsen et al., 2019), Dutch (Boye & Bastiaanse, 2018), Spanish (Martínez-Ferreiro et al., 2019), Western Greenlandic (Nedergaard et al., 2020) and Tagalog (Gerona et al., 2022; Thy et al., 2024) entail the prediction that elements with discursively secondary status will be affected differently across populations (persons with non-fluent aphasia incurring a higher number of errors), and across languages (a given morphosyntactic component may be lexical or discursively primary in one language, and grammatical or discursively secondary in another language, yielding to differential error patterns).

Besides, interpretability and psycholinguistic adequacy of measures should also rely on a theoretical framework for multi-level intervention in treating and assessing discourse macro-structure, as previously outlined by Sherratt (2007), or more recently by the *Linguistic Underpinnings of Narrative in Aphasia* (LUNA[10]) framework proposed by Dipper, Marshall, Boyle, Hersh, et al. (2021) and Dipper et al. (2024). The LUNA framework provides better content and construct validity in relation to discourse and psycholinguistic models. Combined with the LUNA framework, usage-based and functional linguistic approaches offer powerful conceptual tools for describing and interpreting discourse phenomena in aphasia, yielding nuanced cross-linguistic insights into impaired and preserved communicative abilities. Paying close attention to LUNA and usage-based developments in aphasia research will undoubtedly benefit clinicians' practice-based evidence from communication goal-related intervention, enhancing the identification of relevant discourse outcome measures within and across languages.

The implementation of a CLD-COS of measures in relation to communication OMIs

Wallace et al. (2017) further stress the necessity of considering perspectives from PWA and their families in identifying relevant communicative OMIs. This is especially relevant in multilingual/multicultural contexts, where communication needs may involve different languages, code-mixing or code-switching practices, and language-specific therapy goals, which are a relevant aspect of discourse and conversation management, or the willingness to set speech-therapy directions in different languages.

As forcefully argued by Dietz and Boyle (2018b), p. 488, quoting Wallace et al. (2018) statement), "efforts [have to be] directed at standardizing and validating discourse outcome measures (. . .)". They also add that "reference data/norms from NBD speakers, mono- and multilinguals (example, large databases) for each discourse measure are needed". These statements align with our call for continued efforts to identify discourse

outcome measures within and across languages, and to delineate their psychometric properties, including normative dataset development. Moreover, the international consensus study reported by Wallace et al. (2023) (*Research Outcome Measurement in Aphasia*, ROMA-2) sought to establish consensus on outcome measures for assessing functional aspects of communication, incorporating criteria of psychometric quality, clinical relevance, feasibility, and cross-cultural adaptability. Key constructs included language, emotional well-being, communication, patient-reported satisfaction, treatment impact, and quality of life. However, standardized discourse outcome measures, despite their relevance for language and communication, were excluded due to feasibility and standardization challenges. From a preliminary review of relevant publications in the field of communication assessment of PWA, ROMA-2 identified 20 communication measures instruments. Two of them were sets of outcome measures reflecting macro-linguistic discourse analysis: the *Amsterdam-Nijmegen Everyday Language Test* (ANELT, Ruiter et al., 2011, 2022), and a generic category named the "Discourse analysis" which is not a tool, but a series of isolated measures regardless of any protocol (e.g., *Story Grammar measures*; *Utterance/propositional level information measures*; *Correct Information Units*)" (Wallace et al., 2023, p. 1023). The ANELT and the "Discourse analysis" measures were excluded because they did not meet the criterion for face validity against the definition of situated languages, being disconnected from authentic communicative contexts, and considered as non-interactive, non-multimodal, and not relying on common ground.

Nonetheless, connected discourse remains central to aphasia assessment. A key finding of ROMA-2, essential for discourse assessment, is the emphasis on cross-cultural adaptability for OMIs (largest consensus with 74%), affirming the relevance of cross-linguistic applicability.

Wallace et al. (2018) define a Core Outcome Set (COS) as the minimum set of standardized outcomes and measures for clinical trials. Its development involves two phases: first, identifying key outcomes through stakeholder consensus (e.g., patients, clinicians, researchers), then selecting measures and tools for these outcomes *via* a relevant consensus process made possible by international and interdisciplinary collaboration initiatives (Brady et al., 2014). As discussed previously, discourse outcome measures are helpful to assess treatment effects on language and communication. However, due to their lack of standardization and feasibility challenges in aphasia trials, discourse outcome measures were excluded from the first *Research Outcome Measurement in Aphasia* (ROMA) COS iteration (Wallace et al., 2017), despite their clear relevance for evaluating aphasia treatment outcomes comprehensively,[11] and despite encouraging recent results from Dipper et al. (2024), assessing LUNA feasibility in terms of participant recruitment and ongoing involvement, compliance, data completeness, and treatment fidelity.

Nevertheless, to ensure consistent and meaningful assessment of communicative outcomes following treatment using discourse outcome measures, Wallace et al. (2023) advocate for standardizing and validating discourse measures and tools for future inclusion as updates to the existing ROMA-COS iteration.

Conclusion

As a starting point for defining core and cross-linguistic outcome discourse measures, the present reflection underscores key aspects of the linguistic, typological, and

psycholinguistic constructs underlying discourse analysis. Recommendations for a cross-linguistic discourse core outcome set of measures (CLD-COS) should be grounded in comprehensive, theoretically informed, and evidence-based methodological characterizations, particularly concerning reliability and construct validity.

Although some measures may be directly transferable from one language to another, morphosyntactic complexity requires cautious treatment, as language-specific features significantly impact micro-linguistic levels in aphasia. Consequently, guidelines should include a detailed list of core discourse outcome measures, covering micro- and macro-linguistic discourse levels, within and across languages, incorporating the "equivalence principle". Given the ample typological variability that distinguishes natural languages, identifying consistent cross-linguistic features that align specific linguistic impairment patterns with behavioural profiles remains a significant challenge.

Additionally, future practice surveys should also adopt a cross-linguistic lens, recognizing the clinical implications for assessing multilingual PWA. For instance, a practice survey dealing with discourse outcome measures should systematically include targeted questions about the potential cross-linguistic validity of core metrics and solicit input from researchers and clinicians on the challenges and needs involved in assessing discourse production in multiple languages. Another central issue, as highlighted in recent research, is the selection of a limited number of measures deemed most effective in capturing the nature and severity of language deficits, which are often multi-faceted and may vary across linguistic levels in PWA.

Moreover, a central challenge lies in designing an open, flexible annotation method for discourse analysis that accommodates language-specific phenomena without constraining interpretive variability. To ensure reliability and validity in both diagnostic and therapeutic applications across languages, it is critical to establish standardized procedures for scoring and evaluating the psychometric properties of discourse abilities, enhance comparability, and ultimately improve clinical interventions, in connection with existing communication and quality of life OMIs. Building on this perspective, a D-COS and CLD-COS should be developed in alignment with ROMA initiatives, incorporating both structured instruments (such as protocols) and independent discourse measures. One of the main concerns, of course, in this case, will be the anchoring to specific theoretical frameworks rather than merely on individual measures. As Armstrong and Hersh (2024, p. 362) emphasize: "The consensus approach to outcome measures (e.g., as proposed by Dietz & Boyle, 2018b) may be best served through a focus on which theoretical approaches to use (see overview by Linnik et al., 2016) rather than which individual 'measures' to employ – something for future discussion!". To contribute to this crucial in-ground discussion connected to cross-linguistic issues to analyse discourse from monolingual or multilingual PWA, and to develop assessment methods, recent advances driven by a usage-based framework are proving useful towards more comparative, theoretically anchored, and empirically supported approaches. This orientation is expected to contribute meaningfully to the selection of core outcome measures for a CLD-COS. Finally, it is increasingly apparent that a theoretical shift may be required – one that embraces more flexible and transversal functional categories, particularly at the micro-linguistic level of discourse and its morphosyntactic components, to ensure compatibility with the objectives of a multi-level and comprehensive CLD-COS.

Notes

1. COSMIN methodology outlines a four-step process. First, conceptual considerations must be thoroughly examined. Second, existing instruments should be identified through systematic reviews or literature searches. Third, the quality of measures must be evaluated based on their measurement properties and feasibility, following consensus on taxonomy, terminology, and definitions established by Mokkink et al. (2010, see also: https://www.cosmin.nl/tools/cosmin-taxonomy-measurement-properties/). Finally, general recommendations for selecting the most appropriate cross-linguistic discourse measures for a CLD-COS should be formulated.
2. Notice, however, that this hypothesis has been contested by Bastiaanse et al. (2011), who argue that verb processing deficits are more a matter of function than of morphological paradigm complexity.
3. The items concerning cross-linguistic discourse post-test scoring are numbered 514–539 in the BAT manual.
4. AphasiaBank (https://www.talkbank.org/) is part of the *Common Language Resources and Technology Infrastructure* consortium (CLARIN, 2019, https://www.clarin.eu/).
5. The MSSG yields six quantified variables: main concept composite, sequencing, main concept + sequencing, essential story grammar components, total episodic components, and episodic complexity.
6. Stability of discourse measures used or tested in research is also demonstrated by the accuracy of research findings reporting, following recommendations such as best practice guidelines for reporting information about a study design and method (e.g., about participants, sample collection conditions and tasks, discourse coding, rater agreement, analyses and annotations, as suggested by Stark et al., 2022).
7. Frost et al. (2007) and Mokkink et al. (2010) pointed out considerations for evaluating and reporting reliability and validity of patient-reported outcome measures. Reliability refers to the extent to which a measure yields the same number or score each time it is administered (when the construct to be measured has not changed), for the same patient (internal consistency), over time (test-retest reliability), and for the same or different investigator or annotator (intra-rater, inter-rater agreement). Validity refers to the extent to which a measure accurately reflects what it was designed to reflect, rather than something else: the measure is supposed to be defined by one construct and not overlap with other distinct concepts. Subtypes of validity are content, construct, and criterion validity.
8. According to the COSMIN standards (Mokkink, 2010, see also: https://www.cosmin.nl/tools/cosmin-taxonomy-measurement-properties/), the construct validity of a measure is satisfactory if it shows significant differences between NBD speakers and PWA with $p < 0.001/p < 0.01$, for structural validity and hypothesis testing. All discourse measures in different languages should therefore meet this standard.
9. Regarding the intra-rater reliability reflecting the stability criteria, Pritchard et al. (2018) recommend, for English, an intraclass correlation coefficient (ICC) higher than 0.80.
10. LUNA identifies four skills and construct components, each associated with specific measures. The first component is pragmatic, assessed using global informativeness measures. The second component is macro-structure and planning, assessed using story grammar measures to evaluate macro-structural coherence, topic-gist coherence, and informativeness. The third component is propositional, assessed using complete sentences or utterance-related measures to evaluate local coherence, cohesion, informativeness, and semantic-conceptual content. The fourth component is linguistic, focusing on cohesion, informativeness, syntax, lexical-semantic content, and lexical forms using CIUs. Moreover, its use for multi-level treatment, with special focus on everyday life spoken discourse, is based on personal narratives chosen by PWA, and integrates familiar treatments such as the Semantic Feature Analysis.

11. Within the ROMA project, tools such as the *Western Aphasia Battery-Revised* (WAB-R; 74% consensus), the *General Health Questionnaire-12* (GHQ-12; 83%), and the *Stroke and Aphasia Quality of Life Scale* (SAQOL-39; 96%) are recommended as functional, quality-of-life, and communication COS tools for aphasia trials.

Acknowledgements

The authors would like to express their sincere gratitude to the two anonymous reviewers for all their insightful comments, which greatly contributed to improving the manuscript. HS acknowledges support of RIPEC C3 – CNU 07 and University of Toulouse. SMF acknowledges support of RYC2020-028927-1, MICIU/AEI/10.13039/501100011033 and ESF investing in your future.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Halima Sahraoui http://orcid.org/0000-0001-7508-004X
Silvia Martínez-Ferreiro http://orcid.org/0000-0003-2393-1214
Eva Soroli http://orcid.org/0000-0003-2747-9368

References

Ali, M., Soroli, E., Jesus, L. M. T., Cruice, M., Isaksen, J., Visch-Brink, E., Grohmann, K. K., Jagoe, C., Kukkonen, T., Varlokosta, S., Hernandez-Sacristan, C., Rosell-Clari, V., Palmer, R., Martínez-Ferreiro, S., Godecke, E., Wallace, S. J., McMenamin, R., Copland, D.Breitenstein, C. ... Collaboration of Aphasia Trialists (CATs). (2022). An aphasia research agenda - a consensus statement from the collaboration of aphasia trialists. *Aphasiology*, *36*(4), 555–574. https://doi.org/10.1080/02687038.2021.1957081

Alyahya, R. S. W. (2024). The development of a novel, standardized, norm-referenced Arabic discourse assessment tool (ADAT), including an examination of psychometric properties of discourse measures in aphasia. *International Journal of Language & Communication Disorders*, *59*(5), 1460–6984. https://doi.org/10.1111/1460-6984.13083

Arantzeta, M., Howard, D., Webster, J., Laka, I., Martínez-Zabaleta, M., & Bastiaanse, R. (2019). Bilingual aphasia: Assessing cross-linguistic asymmetries and bilingual advantage in sentence comprehension deficits. *Cortex*, *119*, 195–214. https://doi.org/10.1016/j.cortex.2019.04.003

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

Armstrong, E. (2018). The challenges of consensus and validity in establishing core outcome sets. *Aphasiology*, *32*(4), 465–468. https://doi.org/10.1080/02687038.2017.1398804

Armstrong, E., & Hersh, D. (2024). Speaking up and being heard: The importance of functional communication and discourse principles in aphasia intervention. *Seminars in Speech and Language*, *45*(4), 356–367. https://doi.org/10.1055/s-0044-1788981

Arslan, S., & Felser, C. (2018). Comprehension of wh-questions in Turkish-German bilinguals with aphasia: A dual-case study. *Clinical Linguistics and Phonetics*, *32*(7), 640–660. https://doi.org/10.1080/02699206.2017.1416493

Bastiaanse, R. (1995). Broca's aphasia: A syntactic and/or a morphological disorder? A case study. *Brain and Language*, *48*(1), 1–32. https://doi.org/10.1006/brln.1995.1001

Bastiaanse, R., Bamyaci, E., Hsu, C.-J., Lee, J., Duman, T. Y., & Thompson, C. K. (2011). Time reference in agrammatic aphasia: A cross-linguistic study. *Journal of Neurolinguistics*, *24*(6), 652–673. https://doi.org/10.1016/j.jneuroling.2011.07.001

Bastiaanse, R., Jonkers, R., & Thompson, C. K. (2008). *Test for assessing reference of time (TART)*. University of Groningen.

Bates, E., Friederici, A., & Wulfeck, B. (1987). Grammatical morphology in aphasia: Evidence from three languages. *Cortex*, *23*(4), 545–574. https://doi.org/10.1016/S0010-9452(87)80049-7

Bates, E., Wulfeck, B., & MacWhinney, B. (1991). Cross-linguistic research in aphasia: An overview. *Brain and Language*, *41*(2), 123–148. https://doi.org/10.1016/0093-934x(91)90149-u

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Lawrence Erlbaum Associates.

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E., & Schwartz, M. (2000). *Quantitative production analysis: A training manual for the analysis of aphasic sentence production*. Psychology Press.

Bezy, C., Renard, A., & Pariente, J. (2016). *Grémots. Evaluation du Langage dans les Pathologies Neurodégénératives*. Solal.

Bird, H., & Franklin, S. (1996). Cinderella revisited: A comparison of fluent and non-fluent aphasic speech. *Journal of Neurolinguistics*, *9*(3), 187–206. https://doi.org/10.1016/0911-6044(96)00006-1

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

Boxum, E., Van der Scheer, F., & Zwaga, M. (2013). *Analyse voor Spontane Taal bij Afasie (ASTA); Standaard in Samenspraak met de Vereniging voor Klinische Linguïstiek*. 4th version. Retrieved August 11, 2024, from https://klinischelinguistiek.nl/uploads/201307asta4eversie.pdf

Boye, K., & Bastiaanse, R. (2018). Grammatical *versus* lexical words in theory and aphasia: Integrating linguistics and neurolinguistics. *Glossa: A Journal of General Linguistics*, *3*(1), 29, 1–18. https://doi.org/10.5334/gjgl.436

Boye, K., Bastiaanse, R., Harder, P., & Martínez-Ferreiro, S. (2023). Agrammatism in a usage-based theory of grammatical status: Impaired combinatorics, compensatory prioritization, or both? *Journal of Neurolinguistics*, *65*, 101–108. https://doi.org/10.1016/j.jneuroling.2022.101108

Boye, K., & Harder, P. (2012). A usage-based theory of grammatical status and grammaticalization. *Language*, *88*(1), 1–44. https://doi.org/10.1353/lan.2012.0020

Boye, K., Ishkhanyan, B., Theilgaard Brink, E., & Sahraoui, H. (2015). Pronouns and agrammatism in a functional theory of grammatical status. In M. Kohlberger & A. Kloekhorst (Eds.), *SLE 2015 Book of abstracts, 48th Annual meeting of the Societas Linguistica Europaea* (pp. 31–32). Leiden University Centre for Linguistics. https://societaslinguistica.eu/wp-content/uploads/2020/03/SLE2015_BoA.pdf

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

Brady, M. C., Ali, M., Fyndanis, C., Kambanaros, M., Grohmann, K. K., Laska, A.-C., Hernández-Sacristán, C., & Varlokosta, S. (2014). Time for a step change? Improving the efficiency, relevance, reliability, validity and transparency of aphasia rehabilitation research through core outcome measures, a common data set and improved reporting criteria. *Aphasiology*, *28*(11), 1385–1392. https://doi.org/10.1080/02687038.2014.930261

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics and 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

Capilouto, G., Wright, H. H., & Wagovich, S. A. (2005). CIU and main event analyses of the structured discourse of older and younger adults. *Journal of Communication Disorders*, *38*(6), 431–444. https://doi.org/10.1016/j.jcomdis.2005.03.005

CLARIN. (2019, Septembre 1). *European research Infrastructure for language Resources and Technology*. https://www.clarin.eu/

Colin, C., Le Meur, C., Sahraoui, H., & Labrunée, K. (2016). *Adaptation of the AphasiaBank protocol to French language and automated analysis* [Unpublished Thesis Report]. University of Toulouse. https://talkbank.org/aphasia/access/French/Protocol/ColinLeMeur.html

Conner, P. S., Goral, M., Anema, I., Borodkin, K., Haendler, Y., Knoph, M., Mustelier, C., Paluska, E., Melnikova, Y., & Moeyaert, M. (2018). The role of language proficiency and linguistic distance in cross-linguistic treatment effects in aphasia. *Clinical Linguistics and Phonetics*, *32*(8), 739–757. https://doi.org/10.1080/02699206.2018.1435723

Consten, M., & Loll, A. (2012). Circularity effects in corpus studies. Why annotations sometimes go round in circles. *Language Sciences*, *34*(6), 702–714. https://doi.org/10.1016/j.langsci.2012.04.010

Cordella, C., DiFilippo, L., Kolachalama, V. B., & Kiran, S. (2024). Connected speech fluency in poststroke and progressive aphasia: A scoping review of quantitative approaches and features. *American Journal of Speech-Language Pathology*, *33*(4), 20832120. https://doi.org/10.1044/2024_AJSLP-23-00208

Crago, M., Paradis, J., & Menn, L. (2008). Cross-linguistic perspectives on the syntax and semantics of language disorders. In M. J. Ball, M. R. Perkins, N. Müller, & S. Howard (Eds.), *The handbook of clinical linguistics* (pp. 275–289). John Wiley. https://doi.org/10.1002/9781444301007.ch17

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M., & Dipper, L. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language & Communication Disorders*, *55*(3), 417442. https://doi.org/10.1111/1460-6984.12528

Cunningham, K. T., & Haley, K. L. (2020). Measuring lexical diversity for discourse analysis in aphasia: Moving-average type-token ratio and word information measure. *Journal of Speech, Language, & Hearing Research*, *63*(3), 710–721. https://doi.org/10.1044/2019_JSLHR-19-00226

Dalton, S. G. H., Hubbard, H. I., & Richardson, J. D. (2020). Moving toward non-transcription-based discourse analysis in stable and progressive aphasia. *Seminars in Speech and Language*, *41*(1), 32–44. https://doi.org/10.1055/s-0039-3400990

Dalton, S. G., & Richardson, J. D. (2015). Core-lexicon and main-concept production during picture-sequence description in adults without brain damage and adults with aphasia. *American Journal of Speech-Language Pathology*, *24*(4), S923–S938. https://doi.org/10.1044/2015_AJSLP-14-0161

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

Dietz, A., & Boyle, M. (2018b). Discourse measurement in aphasia: Consensus and caveats. *Aphasiology*, *32*(4), 487–492. https://doi.org/10.1080/02687038.2017.1398814

Dik, S. C. (1997). *The theory of functional grammar. Part, 1: The structure of the clause. Part, 2: Complex and derived constructions* (2nd ed.). Mouton de Gruyter.

Dipper, L., Devane, N., Barnard, R., Botting, N., Boyle, M., Cockayne, L., Hersh, D., Magdalani, C., Marshall, J., Swinburn, K., & Cruice, M. (2024). A feasibility randomised waitlist-controlled trial of a personalised multi-level language treatment for people with aphasia: The remote LUNA study. *PLOS ONE*, *19*(6), e0304385. https://doi.org/10.1371/journal.pone.0304385

Dipper, L., Marshall, J., Boyle, M., Botting, N., Hersh, D., Pritchard, M., & Cruice, M. (2021). Treatment for improving discourse in aphasia: A systematic review and synthesis of the evidence base. *Aphasiology*, *35*(9), 1125–1167. https://doi.org/10.1080/02687038.2020.1765305

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

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology*, *25*(11), 14141430. https://doi.org/10.1080/02687038.2011.603898

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, *22*(2), 397–408. https://doi.org/10.1044/1058-0360(2013/12-0083)

Fontan, L., Prince, T., Nowakowska, A., Sahraoui, H., & Martínez-Ferreiro, S. (2023). Automatically measuring speech fluency in people with aphasia: First achievements using read-speech data. *Aphasiology*, *38*(5), 939–956. https://doi.org/10.1080/02687038.2023.2244728

Forbes, M. M., Fromm, D., Holland, A., & MacWhinney, B. (2014). EVAL: A computerized language analysis program for clinicians. In *Clinical Aphasiology Conference*, St. Simons Island. Clinical Aphasiology University of Pittsburgh Library Archive. https://aphasiology.pitt.edu/2572/

Forbes, M. M., Fromm, D., & MacWhinney, B. (2012). AphasiaBank: A resource for clinicians. *Seminars in Speech and Language*, *33*(3), 217–222. https://doi.org/10.1055/s-0032-1320041

Fromm, D., Forbes, M., Holland, A., & MacWhinney, B. (2020). Using AphasiaBank for discourse assessment. *Seminars in Speech and Language*, *41*(1), 10–19. https://doi.org/10.1055/s-0039-3399499

Fromm, D., Katta, S., Paccione, M., Hecht, S., Greenhouse, J., MacWhinney, B., & Schnur, T. T. (2021). A comparison of manual versus automated quantitative production analysis of connected speech. *Journal of Speech, Language, & Hearing Research*, *64*(4), 1271–1282. https://doi.org/10.1044/2020_JSLHR-20-00561

Fromm, D., Macwhinney, B., & Thompson, C. (2020). Automation of the northwestern narrative language analysis system. *Journal of Speech, Language, & Hearing Research*, *63*(6), 1–10. https://doi.org/10.1044/2020_JSLHR-19-00267

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., & Hays, R. D. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health*, *10*, S94S105. https://doi.org/10.1111/j.1524-4733.2007.00272.x

Fyndanis, V., Arcara, G., Christidou, P., & Caplan, D. (2018). Morphosyntactic production and verbal working memory: Evidence from Greek aphasia and healthy aging. *Journal of Speech, Language, & Hearing Research*, *61*(5), 1171–1187. https://doi.org/10.1044/2018_JSLHR-L-17-0103

Gerona, J., Boye, K., Martínez-Ferreiro, S., & Popov, S. (2022). Deictic and anaphoric reference production in tagalog fluent and non-fluent aphasia. *Stem-, Spraak- en Taalpathologie*, *27*, 180–183. https://doi.org/10.21827/32.8310/2022-SA

Goldberg, S., Haley, K. L., & Jacks, A. (2012). Script training and generalization for people with aphasia. *American Journal of Speech-Language Pathology*, *21*(3), 222238. https://doi.org/10.1044/1058-0360(2012/11-0056)

Goodglass, H., & Kaplan, E. (1972). *The assessment of aphasia and related disorders*. Lea & Febiger.

Goral, M., Norvik, M. I., Antfolk, J., Agrotou, I., & Lehtonen, M. (2023). Cross-language generalization of language treatment in multilingual people with post-stroke aphasia: A meta-analysis. *Brain and Language*, *246*, 105326. https://doi.org/10.1016/j.bandl.2023.105326

Gordon, J. K. (1998). The fluency dimension in aphasia. *Aphasiology*, *12*(78), 673–688. https://doi.org/10.1080/02687039808249565

Gordon, J. K. (2006). A quantitative production analysis of picture description. *Aphasiology*, *20*(2–4), 188–204. https://doi.org/10.1080/02687030500472777

Gordon, J. K. (2008). Measuring the lexical semantics of picture description in aphasia. *Aphasiology*, *22*(7-8), 839–852. https://doi.org/10.1080/02687030701820063

Gordon, J. K., & Clough, S. (2024). The flu-ID: A new evidence-based method of assessing fluency in aphasia. *American Journal of speech-Language Pathology*, *33*(6), 2972–2990. https://doi.org/10.1044/2024_AJSLP-23-00424

Halliday, M. A. K. (1985). *An introduction to functional grammar*. Edward Arnold.

Hameister, I., & Nickels, L. (2018). The cat in the tree - using picture descriptions to inform our understanding of conceptualisation in aphasia. *Language, Cognition and Neuroscience*, *33*(10), 1296–1314. https://doi.org/10.1080/23273798.2018.1497801

Hickmann, M. (2003). *Children's discourse: Person, space and time across languages*. Cambridge University Press.

Hickmann, M., & Soroli, E. (2015). From language acquisition to language pathology: Cross-linguistic perspectives. In C. Astésano & M. Jucla (Eds.), *Neuropsycholinguistic perspectives on language cognition. Essays in honor of Jean-Luc Nespoulous* (pp. 46–60). Psychology Press.

Hofstede, B. T. M., & Kolk, H. H. J. (1994). The effects of task variation on the production of grammatical morphology in Broca's aphasia: A multiple case study. *Brain and Language*, *46*(2), 278–328. https://doi.org/10.1006/brln.1994.1017

Holland, A., Forbes, M., Fromm, D., & MacWhinney, B. (2019). Communicative strengths in severe aphasia: The famous people protocol and its value in planning treatment. *American Journal of Speech-Language Pathology*, *28*(3), 1010–1018. https://doi.org/10.1044/2019_AJSLP-18-0283

Holland, A. L. (2021). The value of "communication strategies" in the treatment of aphasia. *Aphasiology*, *35*(7), 984994. https://doi.org/10.1080/02687038.2020.1752908

Huber, W., Poeck, K., & Willmes, K. (1983). *The Aachen Aphasia Test (AAT)*. Hogrefe.

Kertesz, A. (2006). *Western aphasia Battery - Revised*. The Psychological Corporation.

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., & Wright, H. H. (2018). Discourse measurement in aphasia research. *Aphasiology*, *32*(4), 472–474. https://doi.org/10.1080/02687038.2017.1398807

Kolk, H. H. J. (2001). Does agrammatic speech constitute a regression to child language? A three-way comparison between agrammatic, child and normal ellipsis. *Brain and Language*, *77*(3), 340–350. https://doi.org/10.1006/brln.2000.2406

Kolk, H. H. J. (2006). How language adapts to the brain: An analysis of agrammatic aphasia. In L. Progovac, K. Paesani, E. Casielles, & E. Barton (Eds.), *The syntax of nonsententials : Multidisciplinary perspectives* (pp. 229–258). John Benjamins.

Kolk, H. H. J. (2007). Variability is the hallmark of aphasic behaviour: Grammatical behaviour is no exception. *Brain and Language*, *101*(2), 99–102. https://doi.org/10.1016/j.bandl.2007.04.002

Kurland, J., Liu, A., & Stokes, P. (2023). Phase I test development for a brief assessment of transactional success in aphasia: Methods and preliminary findings of main concepts in non-aphasic participants. *Aphasiology*, *37*(1), 39–68. https://doi.org/10.1080/02687038.2021.1988046

Kuzmina, E., Goral, M., Norvik, M., & Weekes, B. S. (2019). What influences language impairment in bilingual aphasia? A meta-analytic review. *Frontiers in Psychology*, *10*, 445. https://doi.org/10.3389/fpsyg.2019.00445

Lavoie, M., Black, S. E., Tang-Wai, D. F., Graham, N. L., Stewart, S., Freedman, M., Leonard, C., & Rochon, E. (2022). Longitudinal changes in connected speech over a one-year span in the nonfluent/agrammatic variant of primary progressive aphasia. *Aphasiology*, *37*(8), 1186–1197. https://doi.org/10.1080/02687038.2022.2084707

Lavoie, M., Black, S. E., Tang-Wai, D. F., Graham, N. L., Stewart, S., Leonard, C., & Rochon, E. (2021). Description of connected speech across different elicitation tasks in the logopenic variant of primary progressive aphasia. *International Journal of Language & Communication Disorders*, *56*(5), 1074–1085. https://doi.org/10.1111/1460-6984.12660

Leaman, M. C., & Edmonds, L. A. (2021a). Assessing language in unstructured conversation in people with aphasia: Methods, psychometric integrity, normative data, and comparison to a structured narrative task. *Journal of Speech, Language, & Hearing Research*, *64*(11), 4344–4365. https://doi.org/10.1044/2021_JSLHR-20-00641

Leaman, M. C., & Edmonds, L. A. (2021b). Measuring global coherence in people with aphasia during unstructured conversation. *American Journal of Speech-Language Pathology*, *30*(1), 359–375. https://doi.org/10.1044/2020_AJSLP-19-00104

Leaman, M. C., & Edmonds, L. A. (2023). Analyzing language in the picnic scene picture and in conversation: The type of discourse sample we choose influences findings in people with aphasia. *American Journal of Speech-Language Pathology*, *32*(4), 1413–1430. https://doi.org/10.1044/2023_AJSLP-22-00279

Lecours, A. R., Nespoulous, J.-L., & Joanette, Y. (1996). *Protocole Montréal-Toulouse (MT86)*. Ortho Edition.

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

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, & Hearing Research*, *66*(7), 2421–2433. https://doi.org/10.1044/2023_JSLHR-22-00642

Liu, J., Wumaier, A., Fan, C., & Guo, S. (2023). Automatic fluency assessment method for spontaneous speech without reference text. *Electronics*, *12*(8), 1775. https://doi.org/10.3390/electronics12081775

MacCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual lexical diversity (MTLD)* [Unpublished PhD Dissertation]. University of Memphis.

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk. Part 1: The CHAT transcription format. Part 2: The CLAN program* (3rd ed.). Lawrence Erlbaum Associates.

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

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

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

Martínez-Ferreiro, S., & Bastiaanse, R. (2013). Time reference in Spanish and Catalan non-fluent aphasia. *Lingua*, *137*, 88–105. https://doi.org/10.1016/j.lingua.2013.09.003

Martínez-Ferreiro, S., Bastiaanse, R., & Boye, K. (2020). Functional and usage-based approaches to aphasia: The grammatical-lexical distinction and the role of frequency. *Aphasiology*, *34*(8), 927–942. https://doi.org/10.1080/02687038.2019.1615335

Martínez-Ferreiro, S., Ishkhanyan, B., Rosell-Clarí, V., & Boye, K. (2019). Prepositions and pronouns in connected discourse of individuals with aphasia. *Clinical Linguistics and Phonetics*, *33*(6), 497–517. https://doi.org/10.1080/02699206.2018.1551935

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Méndez-Orellana, C., Holme, C., & Martínez-Ferreiro, S. (2022). *Manual del Análisis del Lenguaje Espontáneo ALEA en Personas Adulta*, Patricia Avilés Retamal, (Ed.). Pontificia Universidad Católica de Chile. https://lenguajespontaneo.cl/

Menn, L., & Obler, L. K. (1990a). *Agrammatic aphasia: A cross-language narrative sourcebook*. John Benjamins.

Menn, L., & Obler, L. K. (1990b). Cross-language data and theories of agrammatism, chapter 20. In L. Menn & L. K. Obler (Eds.), *Agrammatic aphasia: A cross-language narrative sourcebook* (Vol. 2, pp. 1369–1389). John Benjamins.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Nedergaard, J. S. K., Martínez-Ferreiro, S., Fortescue, M. D., & Boye, K. (2020). Non-fluent aphasia in a polysynthetic language: Five case studies. *Aphasiology*, *34*(6), 675–694. https://doi.org/10.1080/02687038.2019.1643000

Nespoulous, J.-L. (2000). Invariance vs variability in aphasic performance. An example: Agrammatism. *Brain and Language*, *71*(1), 167–171. https://doi.org/10.1006/brln.1999.2242

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

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

Nielsen, S. R., Boye, K., Bastiaanse, R., & Lange, V. M. (2019). The production of grammatical and lexical determiners in Broca's aphasia. *Language, Cognition and Neuroscience*, *34*(8), 1027–1040. https://doi.org/10.1080/23273798.2019.1616104

Nilipour, R. (2000). Agrammatic language: Two cases from Persian. *Aphasiology*, *14*(12), 1205–1242. https://doi.org/10.1080/02687030050205723

O'Connor, B., Anema, I., Hia, D., Signorelli, T., & Obler, L. K. (2005). Agrammatism: A cross-linguistic clinical perspective. *The ASHA Leader Archive*, *10*(17), 8–29. https://doi.org/10.1044/leader.FTR3.10172005.8

Ossewaarde, R., Jonkers, R., Jalvingh, F., & Bastiaanse, R. (2020). Quantifying the uncertainty of parameters measured in spontaneous speech of speakers with dementia. *Journal of Speech, Language, & Hearing Research*, *63*(7), 2255–2270. https://doi.org/10.1044/2020_JSLHR-19-00222

Paradis, M. (1988). Recent developments in the study of agrammatism: Their import for the assessment of bilingual aphasia. *Journal of Neurolinguistics*, *3*(2), 127–160. https://doi.org/10.1016/0911-6044(88)90012-7

Paradis, M. (2001). The need for awareness of aphasia symptoms in different languages. *Journal of Neurolinguistics*, *14*(2-4), 85–91. https://doi.org/10.1016/S0911-6044(01)00009-4

Paradis, M. (2011). Principles underlying the bilingual aphasia test (BAT) and its uses. *Clinical Linguistics and Phonetics*, *25*(6-7), 427–443. https://doi.org/10.3109/02699206.2011.560326

Paradis, M., & Libben, G. (1987). *The assessment of bilingual aphasia*. Lawrence Erlbaum.

Penn, C., & Beecham, R. (1992). Discourse therapy in multilingual aphasia: A case study. *Clinical Linguistics and Phonetics*, *6*(1-2), 11–25. https://doi.org/10.3109/02699209208985516

Penn, C., Venter, A., & Ogilvy, D. (2001). Aphasia in Afrikaans: A preliminary analysis. *Journal of Neurolinguistics*, *14*(2-4), 111–132. https://doi.org/10.1016/S0911-6044(01)00011-2

Powell, T. W. (2006). A reliability study of BDAE-3 discourse coding. *Clinical Linguistics and Phonetics*, *20*(7-8), 607–612. https://doi.org/10.1080/02699200500266679

Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, *18*(12), 1075–1092. https://doi.org/10.1080/02687030444000534

Prinsen, C. A. C., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P. R., & Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a "core outcome set" - a practical guideline. *Trials*, *17*(1), 449. https://doi.org/10.1186/s13063-016-1555-2

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. (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

Richardson, J. D., Dalton, S. G., Greenslade, K. J., Jacks, A., Haley, K. L., & Adams, J. (2021). Main concept, sequencing, and story grammar analyses of cinderella narratives in a large sample of persons with aphasia. *Brain Sciences*, *11*(1), 110. https://doi.org/10.3390/brainsci11010110

Richardson, J. D., & Dalton, S. G. H. (2020). Main concepts for two picture description tasks: An addition to Richardson and Dalton, 2016. *Aphasiology*, *34*(1), 119–136. https://doi.org/10.1080/02687038.2018.1561417

Rochon, E., Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language*, *72*(3), 193–218. https://doi.org/10.1006/brln.1999.2285

Ruigendijk, E., Van Zonneveld, R., & Bastiaanse, R. Y. R. M. (1999). Case assignment in agrammatism. *Journal of Speech, Language, & Hearing Research*, *42*(4), 962–971. https://doi.org/10.1044/jslhr.4204.962

Ruiter, M. B., Kolk, H. H. J., Rietveld, T. C. M., Dijkstra, N., & Lotgering, E. (2011). Towards a quantitative measure of verbal effectiveness and efficiency in the Amsterdam-Nijmegen everyday language test (ANELT). *Aphasiology*, *25*(8), 961–975. https://doi.org/10.1080/02687038.2011.569892

Ruiter, M. B., Otters, M. C., Piai, V., Lotgering, E. A. M., Theunissen, J. E. M. C., & Rietveld, T. C. M. (2022). A transcription-less quantitative analysis of aphasic discourse elicited with an adapted version of

the Amsterdam-Nijmegen everyday language test (ANELT). *Aphasiology*, *37*(10), 1556–1575. https://doi.org/10.1080/02687038.2022.2109124

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, *37*(3), 440–479. https://doi.org/10.1016/0093-934x(89)90030-8

Sahraoui, H. (2009). *Neuropsycholinguistic study of compensatory strategies in agrammatism: Quantitative and functional characterisation of variability* [Unpublished Ph.D. Dissertation]. University of Toulouse,

Sahraoui, H. (2015). Describing and interpreting variability in agrammatic speech production. In C. Astésano & M. Jucla (Eds.), *Neuropsycholinguistic perspectives on language cognition. Essays in honor of Jean-Luc Nespoulous* (pp. 131–143). Psychology Press.

Sahraoui, H., & Nespoulous, J.-L. (2012). Across-task variability in agrammatic performance. *Aphasiology*, *26*(6), 785–810. https://doi.org/10.1080/02687038.2011.650625

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

Sherratt, S. (2007). Multi-level discourse analysis: A feasible approach. *Aphasiology*, *21*(3-4), 375–393. https://doi.org/10.1080/02687030600911435

Soroli, E. (2011). *Language and spatial cognition in English and in French: Cross-linguistic perspectives in aphasia* [Unpublished Ph.D. Dissertation]. University of Paris 8.

Soroli, E. (2018). Event processing in agrammatic aphasia: Does language guide visual processing and similarity judgments? *Aphasiology*, *32*(1), 219–221. https://doi.org/10.1080/02687038.2018.1489123

Soroli, E. (2024). How language influences spatial thinking, categorization of motion events, and gaze behavior: A cross-linguistic comparison. *Language and Cognition*, *16*(4), 924–968. https://doi.org/10.1017/langcog.2023.66

Soroli, E., Sahraoui, H., & Sacchett, C. (2012). Linguistic encoding of motion events in English and French: Typological constraints on second language acquisition and agrammatic aphasia. *Language, Interaction and Acquisition*, *3*(2), 261–287. https://doi.org/10.1075/lia.3.2.05sor

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, O. D. B., Brock, K., McKinney, B. 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

Strömqvist, S., & Verhoeven, L. (Eds.). (2004). *Relating events in narrative: Typological and contextual perspectives*. Lawrence Erlbaum Associates.

Sung, J. E., De De, G., & Lee, S. E. (2016). Cross-linguistic differences in a picture-description task between Korean- and English-speaking individuals with aphasia. *American Journal of speech-Language Pathology*, *25*(4S). https://doi.org/10.1044/2016_AJSLP-15-0140

Swinburn, K., Porter, G., & Howard, D. (2012). *Comprehensive aphasia test (CAT)*. Psychology Press.

Talmy, L. (2000). *Toward a cognitive semantics. Volume, 1: Concept structuring systems. Volume, 2: Typology and process in concept structuring*. MIT Press.

Thompson, C. K., Meltzer-Asscher, A., Cho, S., Lee, J., Wieneke, C., Weintraub, S., & Mesulam, M.-M. (2013). Syntactic and morphosyntactic processing in stroke-induced and primary progressive aphasia. *Behavioural Neurology*, *26*(1–2), 35–54. https://doi.org/10.1155/2013/749412

Thy, A. D. M., Gerona, J., Martínez-Ferreiro, S., Popov, S., & Boye, K. (2024). Deictic vs. anaphoric pronouns: A comparison of fluent and non-fluent aphasia in English and tagalog. *Language, Cognition and Neuroscience*, *39*(7), 1–15. https://doi.org/10.1080/23273798.2024.2368114

Van der Scheer, F., Zwaga, M., & Jonkers, R. (2011). Normering van de ASTA, Analyse voor Spontane Taal bij Afasie. *Stem-, Spraak- en Taalpathologie*, *17*(2), 19–30.

Van Dijk, T. (1997). The study of discourse. In T. Van Dijk (Ed.), *Discourse as structure and process. Discourse studies: A multidisciplinary introduction* (Vol. 1, pp. 1–34). SAGE. https://doi.org/10.4135/9781446221884

Wallace, S. J., Worrall, L. E., Rose, T., & Le Dorze, G. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? A core outcome set . . . or greater standardisation of discourse measures? *Aphasiology*, *32*(4), 479–482. https://doi.org/10.1080/02687038.2017.1398811

Wallace, S. J., Worrall, L., Rose, T. A., Alyahya, R. S. W., Babbitt, E., Beeke, S., de Beer, C., Bose, A., Bowen, A., Brady, M., Breitenstein, C., Bruehl, S., Bryant, L., Cheng, B. B. Y., Cherney, L. R., Conroy, P., Copland, D. A., Croteau, C.Cruice, M. . . . Le Dorze, G. (2023). Measuring communication as a core outcome in aphasia trials: Results of the ROMA-2 international core outcome set development meeting. *International Journal of Language & Communication Disorders*, *58*(4), 1017–1028. https://doi.org/10.1111/1460-6984.12840

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Cruice, M., Isaksen, J., Kong, A. P. H., Simmons-Mackie, N., Scarinci, N., & Gauvreau, C. A. (2017). Which outcomes are most important to people with aphasia and their families? An international nominal group technique study framed within the ICF. *Disability and Rehabilitation*, *39*(14), 1364–1379. https://doi.org/10.1080/09638288.2016.1194899

Wechsler, D. (1981). *Manual for the Wechsler adult intelligence scale - Revised*. Psychological Corporation.

Whitworth, A., Claessen, M., Leitão, S., & Webster, J. (2015). Beyond narrative: Is there an implicit structure to the way in which adults organise their discourse? *Clinical Linguistics and Phonetics*, *29*(6), 455–481. https://doi.org/10.3109/02699206.2015.1020450

Wright, H. H., & Capilouto, G. J. (2012). Considering a multi-level approach to understanding maintenance of global coherence in adults with aphasia. *Aphasiology*, *26*(5), 656–672. https://doi.org/10.1080/02687038.2012.676855

Wright, H. H., Koutsoftas, A., Fergadiotis, G., & Capilouto, G. (2010). Coherence in stories told by adults with aphasia. *Procedia - Social & Behavioral Sciences*, *6*, 111–112. https://doi.org/10.1016/j.sbspro.2010.08.056

Zanini, S., Tavano, A., & Fabbro, F. (2010). Spontaneous language production in bilingual Parkinson's disease patients: Evidence of greater phonological, morphological and syntactic impairments in native language. *Brain and Language*, *113*(2), 84–89. https://doi.org/10.1016/j.bandl.2010.01.005

Appendix A Set of quantitative measures most commonly used, adapted and refined in research, taken from key cross-linguistic discourse analysis protocols: BAT, QPA, EVAL, related to discourse domains (productivity, information content and grammatical complexity)

3 types of speech and discourse variables and their subtypes (536 discourse measures listed, taken from Bryant et al., 2016, pp. 517-518)	*Bilingual Aphasia Test* **(BAT)** (Paradis & Libben, 1987, pp. 191-192: post-test analysis of part B - Spoken discourse) -All languages -Spontaneous speech - interview (measures 514 to 539) & descriptive speech - storytelling: "the nest story" (6 pictures) (measures 540 to 565)	*Quantitative Production Analysis* **protocol (QPA)** (Rochon et al., 2000, p. 201) -Original measures for English corpora -Narrative discourse	*EVAL* (Forbes et al., 2012; MacWhinney, 2000, pp. 132-138) -Original measures for English corpora -All connected discourse types

(*Continued*)

(Continued).

Verbal productivity	**Sample length** - quantification of the amount of language units within a sample (e.g., number of words, clauses, sentences, etc.)	**(2 ratios)** -Number of utterances -Total number of words (5 minutes sample) -MLUa: Mean length of utterance -MLUb: Mean length of utterance of the 5 longest utterances -The discourse is cohesive (yes: +; no: 0) -The discourse is pragmatically sound (yes: +; no: 0)		-Duration: total time of the sample in hours:minutes: seconds -Total Utts: total utterances -MLU Utts: number of utterances used to compute MLU
	Lexical diversity - the variety of vocabulary used in a sample (e.g., type-token ratio (TTR), number of different words (NDW), vocabulary diversity (D), etc.)	**(1 ratio)** -Number of different words -Type/token ratio *(lexical, exclusion of free standing and inflectional morphemes)*		**(1 ratio)** -FREQ TTR: type/token ratio
	Speech fluency - the rate of speech production in units per minute, or measurement of the length or number of dysfluent moment (e.g., words per minute (WPM), proportion of dysfluencies per word, etc.)	-Number of intraphrasal pauses	**(2 ratios)** -Speech rate (WPM) -Proportion of narrative words/ total words	**(1 ratio)** -Speech rate (WPM)
	Word finding behaviours - linguistic structures or behaviours that indicate difficulty retrieving words (e.g., paraphasias, delays, attempts at word production, etc.)	-Number of inappropriate foreign words -Number of neologisms -Number of phonemic paraphasias resulting in nonwords -Number of phonemic paraphasias resulting in words -Number of semantic paraphasias -Number of verbal paraphasias -Number of perseverations -Number of circumlocutions -Number of stereotypic phrases -Evidence of word-finding -difficulty -Detection of foreign accent (0 to 5-0: none; 5: very strong)		**(1 ratio)** -% Word Errors: percentage of words that are coded as errors -Utt Errors: number of utterances coded as errors -Retracing [//]: number retracings (self-corrections or changes) -Repetition [/]: number of repetitions

(*Continued*)

(Continued).

Information content	**Efficiency** - the rate of information production per minute (e.g., Correct Information Units per minute (CIUs/min), etc.)		
	Lexical - analysis of single words that are integral to the expression of the information content of the sample as a whole (e.g., correct information units (CIUs), propositional density, number of content words, etc.)		**(1 ratio)** -Density: measure of propositional idea density
	Semantic/conceptual - the elements that communicate the gist, theme or main ideas of the language sample (e.g., information units (IUs), main events, story propositions, etc.)	-Number of individual sentences that are semantically deviant	
	Schema related - analysis of the over-arching schematic structure or frame work, specific to the elicited language genre, onto which information content is mapped (e.g., number of temporal-causal sequences, utterances conveying background or setting, number of procedural steps communicated, etc.)		
	Cohesion - the linguistic structures that tie the smaller lexical and grammatical elements of a language sample together to form a whole (e.g., number of cohesive ties, number of referential errors, use of conjunctions, etc.)		

(*Continued*)

(Continued).

Grammatical complexity	**Morphological** - the grammatical structure of language measured in the use of bound and unbound morphemes (e.g., number of bound morphemes, use of tense or plurals, etc.)	-Number of paragrammatisms -Number of missing obligatory grammatical morphemes *(free standing + inflectional: all in a single category)*	**(2 ratios)** -Proportion of verb inflected/possible inflections (Inflection Index) -Elaboration of auxiliary (Aux score)	**(6 ratios)** % Plurals, % 3S, % 1S/3S, % Past, % PastP: past participle, % PresP: present participle
	Word classes - the grammatical categories of words used within a language sample (e.g., numbers of nouns, verbs, adjectives, closed-class words, etc.)		**(4 ratios)** -Proportion closed class words/ narrative words (CC) -Determiner/noun ratio (Determiner Index) -Proportion of pronouns (P/N+P) -Proportion of verbs (V/N+V)	**(11 ratios)** % Nouns, % Verbs, % Aux, % Mod, % prep: prepositions, %adv: adverbs, %adj: adjectives, % conj: conjunctions, % det: determiners, % pro: pronouns, noun/verb ratio: total # of nouns ÷ total # of verbs, open/closed ratio: total # of open class words ÷ total # of closed class words, #open-class: total # of open class words, #closed-class: total # of closed class words
	Syntactic - the structure of language at the sentential level (e.g., proportion of sentences that are grammatically complete, number of words in sentences, etc.)	**(1 ratio)** -Number of word-order errors -Number of verbs per utterance -Number of subordinate clauses	**(5 ratios)** -Proportion of words in sentences -Proportion of well-formed sentences -Structural elaboration of sentences (Sentence elaboration index) -Embedding index -Median length of utterance (MLU)	**(3 ratios)** -Verbs/Utt: verbs per utterance: (roughly corresponds to clauses per utterance) -MLU Words: MLU in words -MLU Morphemes

Appendix B Set of discourse measures retained in norm referenced protocols for discourse assessment in Arabic and French-Canadian, related to discourse domains (productivity, information content and grammatical complexity)

3 types of speech and discourse variables and their subtypes (536 discourse measures listed, from Bryant et al., 2016, pp. 517-518)		Norm referenced protocol for Arabic (*Arabic Discourse Assessment Tool* - ADAT, Alyahya, 2024) -3 tasks: Picture description, Storytelling, Procedural discourse -Using original coding and linguistic analysis	Norm referenced protocol for French-Canadian (Boucher et al., 2022) -1 task: WAB-Picnic Scene -Using CLAN program analyses (MacWhinney, 2000)
Verbal productivity	Sample length	-Token counts -Sample duration in seconds	-Total number of words -Sample duration
	Lexical diversity	-Type-token ratio (TTR) -Number of different words (NDW)	-Lexical diversity (VOC-D measure)
	Speech fluency	-Speech rate (words per minute)	-Speech rate (words per minute)
	Word finding behaviours		-Speech errors (repetitions, self-corrections, and word errors)
Information content	Efficiency		-Communication efficiency: .ICUs/duration (mean number of ICU conveyed per second) .ICUs/token (number of ICUs divided by total number of words) .ICUs/utterance (mean number of ICUs produced per utterance)
	Lexical	-Correct information units (CIU)	
	Semantic/ conceptual	-Lexical information units (LIU)	
	Schema related		
	Cohesion		
Grammatical complexity	Morphological	-Number of morphemes	
	Word classes	-Proportion of nouns and verbs -Proportion of open and closed word classes	-Lexical selection (open-to-closed class ratio and noun-to-verb ratio)
	Syntactic	-Mean length of utterance (MLU) -Number of complete sentences	-Mean length of utterance (MLU) -Syntactic complexity (verbs/utterance)