Received: 30 June 2023 | Accepted: 8 March 2024

DOI: 10.1111/1460-6984.13028

International Journal of Language & Communication Disorders

REVIEW

A scoping review of transcription-less practices for analysis of aphasic discourse and implications for future research

Brielle C. Stark[1,*] | Sarah Grace Dalton[2,*]

[1]Department of Speech, Language and Hearing Sciences, Indiana University Bloomington, Bloomington, Indiana, USA

[2]Department of Speech Pathology and Audiology, Marquette University, Milwaukee, Wisconsin, USA

Correspondence
Brielle C. Stark, Department of Speech, Language and Hearing Sciences, Indiana University Bloomington, IU Health Sciences Building, 2631 East Discovery Parkway, Bloomington, IN 47408, USA.
Email: bcstark@iu.edu

*Brielle C. Stark and Sarah Grace Dalton contributed equally to this work.

Abstract

Background: It is important to capture a comprehensive language profile from speakers with aphasia. One way to do this is to evaluate spoken discourse, which is language beyond a single simple clause used for a specific purpose. While the historical trend in aphasiology has been to capture performance during isolated language tasks, such as confrontation naming, there is a demonstrated need and benefit to collecting language information from tasks that resemble everyday communication. As a result, there has been an increase in discourse analysis research over time. However, despite clinicians' and researchers' desire to analyse spoken discourse, they are faced with critical barriers that inhibit implementation.

Aims: To use scoping review methodology to identify transcription-less tools developed to analyse discourse from individuals with aphasia. The review addressed the following question: 'What transcription-less tools and analysis procedures are available to assess discourse in people with aphasia?' and included several sub-questions to further characterise the type of discourse and tool being used, participants on whom the tool was used to rate discourse abilities, tool users (raters), and psychometric properties.

Methods: The scoping review was conducted between the months of October 2022 and January 2023, concluding 30 January 2023, on PubMed/NCBI, Academic Search Complete and Linguistics and Language Behavior Abstracts. Major inclusion parameters included peer-reviewed papers written in English; that the tool was used to analyse discourse elicited by individuals with acquired aphasia; and that the tool was not a part of a standardised battery or assessment. Perceptual discourse analysis was defined as any analysis which primarily relied on listener impressions and did not numerically quantify specific language behaviours. 'Transcription-less' analysis was defined as any discourse analysis which did not require a written record of the discourse sample in order to be completed. A total of 396 abstracts were screened and 39 full articles were reviewed, yielding 21 papers that were included in the review.

Main Contribution: An overview of the state of transcription-less tools for aphasic discourse analysis is provided, and next steps are identified to facilitate increased implementation of discourse analysis in clinical and research settings.

Conclusion: Transcription-less tools have many benefits for analysing multiple levels (e.g., linguistic, propositional, macrostructural, pragmatic) of discourse, but require more research to establish sound psychometric properties and to explore the implementation of these tools in clinical settings.

KEYWORDS
aphasia, analysis, discourse

What this paper adds

What is already known on this subject

- Individuals with aphasia prioritise treatment outcomes at the discourse level such as being able to engage in conversations with friends and family about important topics and participating in social and leisure activities. However, discourse is rarely used as a treatment outcome measure in clinical practice due to multiple barriers. When speech-language pathologists do assess discourse, they often make perceptual judgements without transcribing the discourse sample. Transcription-less analysis procedures may improve clinical implementation of discourse assessment, which would better match treatment outcome measurement to clients' desired outcomes. However, little is known about the current state of transcription-less discourse analysis, blocking progress.

What this paper adds to existing knowledge

- This study provides an overview of currently available transcription-less discourse analysis procedures that are not part of published standardised aphasia assessments. Transcription-less measures are available to evaluate discourse at all levels (i.e., lexical, propositional, macro-structural/planning, and pragmatic) and most measures include items that assess discourse abilities across multiple levels. Additionally, there are transcription-less measures available for both structured (e.g., picture scene description) and spontaneous (e.g., conversation) discourse tasks. However, current transcription-less procedures are lacking psychometric data including information about validity and reliability.

What are the potential or actual clinical implications of this work?

- Transcription-less analysis methods may provide an avenue for increased implementation of discourse measurement into clinical practice. Further research is needed to determine the clinical utility of transcription-less discourse analysis to better monitor clients' desired treatment outcomes.

INTRODUCTION

Evaluating spoken discourse—language beyond a single simple clause used for a specific purpose (Armstrong, 2000)—has gained considerable traction in aphasiology over the last 30 years (Armstrong, 2000; Bryant et al., 2016; Dietz & Boyle, 2018a, 2018b; Kintz & Wright, 2017; Linnik et al., 2016; Stark, Dutta, Murray, Bryant et al., 2021). Discourse has two forms: monologic and dialogic. In empirical settings, like speech and language assessment settings, discourse is elicited by employing a variety of instructions, usually grouped into genres (e.g., narrative, conversation) and, within those genres, tasks (e.g., fictional versus. personal narrative, interview versus. open conversation). It is best practice to acquire discourse samples across genres and tasks to form a representative sampling of language (Armstrong, 2000; Brookshire & Nicholas, 1994; Leaman & Edmonds, 2023; Stark, 2019; Stark & Fukuyama, 2021; Ulatowska et al., 1981; Wright & Capilouto, 2009). Armstrong (2000) highlights that features (e.g., number of words) extracted from a single task are not generalisable beyond that task, and therefore researchers and clinicians are cautioned about broader interpretation related to treatment effectiveness and recovery of language derived from a single genre or task.

A recent review of the substantial aphasic discourse literature proposed a unified theoretical framework with four building blocks, or levels, of discourse: linguistic, propositional, macrostructural/planning and pragmatic (Dipper et al., 2021). To create this theoretical framework, authors leveraged 10 other frameworks from the fields of linguistics, cognitive linguistics, psycholinguistics, sociolinguistics and pragmatics (Frederiksen et al., 1990; Halliday, 1985; Kintsch & van Dijk, 1978; Labov, 1972; Levelt, 1989; Rumelhart, 1975; Slobin, 1996; Sperber & Wilson, 1986; Stein & Glenn, 1979; van Dijk & Kintsch, 1983). Readers are referred to the Dipper et al., 2021 article for more information about the synthesis of this new theoretical framework from existing frameworks. As explicitly noted by Dipper et al. (2021), theoretical frameworks are imperative in research and clinical practice. The Dipper et al. (2021), theoretical framework is the first created especially with aphasic spoken discourse in mind. That is—the theoretical framework leverages the evidence base, which demonstrates conclusively that aphasia impacts discourse across these four levels (Bryant et al., 2016; Linnik et al., 2016; Pritchard et al., 2017). In clinical practice, these theoretical frameworks can be used to identify intact and impaired components of discourse processing to inform treatment planning. For example, clinicians may want to systematically evaluate the extent to which a therapy improves different levels of discourse. Within each level of discourse, researchers and clinicians measure features (sometimes called 'proxies') that are thought to be representative of a core function of that level. Dipper et al. (2021) give several examples of features that can be measured from each level, such as:

- Linguistic: syntax, lexical semantics, lemma and lexemes, and phonology.
- Propositional: sequencing, sentence semantics, cohesion (grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning) and semantic content.
- Macrostructural/Planning: structure, story content, framing, local or global coherence (respectively, relations between sentences or propositions of a text or across the text which lend it meaning) and gist (involving propositional content as well as local and global coherence).
- Pragmatic: context, interpersonal factors, interactional factors and influences on discourse from situational and external influences.

These levels are useful in conceptualising the type of information being extracted from discourse in published papers, as well as in the clinic. Examples of specific features that measure these constructs seen in the aphasia discourse literature include percentage paraphasias (e.g., Stark et al., 2019) and correct information units (Nicholas & Brookshire, 1993) at the linguistic level; main concept analysis (Dalton & Richardson, 2019; Nicholas & Brookshire, 1995) and sequencing (Richardson et al., 2021) at the propositional level; story grammar (Stein & Glenn, 1975) at the macrostructural/planning level; and the evaluation of context's influence on discourse, for example, how discourse is moulded across different tasks, communication partners, communication purposes or in monologic versus dialogic settings (Doyle et al., 1995; Fergadiotis & Wright, 2011; Stark, 2019; Stark & Fukuyama, 2021; Ulatowska et al., 1981, 1990) at the pragmatic level.

While monologic discourse is often analysed, analysing dialogue also provides a variety of information specific to each discourse level. For example, dialogue can elicit data around informational redundancy, use of nonspecific vocabulary, message accuracy, topic maintenance, response appropriateness, situational appropriateness, revision behaviours and turn-taking difficulty. Frameworks like Conversation Analysis have been useful for analysing dialogue (e.g., Damico et al., 1999). Some of the information extracted from dialogue fits clearly into the four discourse levels. For example, 'topic maintenance' is similar to coherence and thus reflects the macrostructural/planning level; 'use of nonspecific vocabulary' reflects the linguistic level; and 'situational appropriateness' reflects the pragmatic level. However, some behaviours, like 'revision behaviour' and 'turn taking', are not easily parsed into a single discourse level. For example, 'revision behaviour' may reflect aspects of cohesion (propositional level) and/or coherence (macrostructural/planning level). Turn-taking behaviour (or lack of turn taking) may reflect impairments at the macrostructural/planning and/or pragmatic level.

Of note, while Dipper and colleagues' framework is helpful for conceptualising different levels of discourse, it does not directly address non-linguistic (e.g., gesture, eye gaze) and paralinguistic features (e.g., tone of voice), or other behaviours (e.g., speech characteristics like dysarthria, pausing), which are likewise relevant and meaningful in aphasic discourse production. This may be in part because non-verbal, paralinguistic and 'other' behaviours are not easily confined to a single discourse category, or because the traditional focus of discourse analysis in aphasia has been heavily weighted toward transcription, which makes it more difficult to capture these features (Bryant et al., 2016). For example, a gesture tends to span across an utterance rather than a single word, be transient, and may or may not be directly related to the language produced.

This scoping review will draw upon the Dipper et al. (2021), framework in order to contextualise the extent to which transcription-less tools have evaluated discourse in aphasia.

Benefits to analysing aphasic spoken discourse

Multiple factors have driven the increasing focus on discourse assessment and analysis in individuals with aphasia over the past three decades. Individuals with aphasia prioritise communication ability, especially conversational and narrative skills, as a top rehabilitation outcome (Worrall et al., 2011). Unsurprisingly, a large majority of researchers and clinicians want to include discourse analysis in their practice (Bryant et al., 2017; Cruice et al., 2020; Stark, Dutta, Murray, Fromm et al., 2021). Researchers and clinicians agree that assessing and analysing discourse enables comprehensive characterisation of language and its use (Dipper et al., 2021; Marini et al., 2011; Stark, Dutta, Murray, Fromm et al., 2021). For example, it is possible to evaluate linguistic, propositional, macrostructural and pragmatic abilities from a single spoken discourse sample. This is not possible when evaluating isolated language skills, such as naming or repetition ability. Because of the variety of elicitation methods that exist for monologic discourse (e.g., narrative, picture description or exposition), evaluating discourse also enables evaluation of the interaction of language with other cognitive processes, such as executive function and long-term memory, and how language changes when the topic becomes more salient and tellable. A more tellable task likely enhances many aspects of the discourse, such as global coherence (Ulatowska & Bond Chapman, 1989; Ulatowska & Olness, 2004; Ulatowska, Doyel et al., 1983; Ulatowska, Freedman-Stern et al., 1983). Discourse provides a means of understanding how language is influenced by dialect, culture and other important variables, such as ethnicity and geographic location (Olness et al., 2002; Ulatowska et al., 2003). Further, analysing discourse may provide the most sensitive and accurate portrayal of language ability in individuals with the mildest aphasia, who typically test at ceiling on standardised batteries and because of this, are clinically under-served (Fromm et al., 2017).

Eliciting discourse enables evaluation of multimodal communication components, such as manual gesture, which is known to be more prevalent in people with aphasia and used to supplement speech (de Beer et al., 2019, 2020; Lanyon & Rose, 2009; Sekine & Rose, 2013; van Nispen et al., 2017). Lastly, characterising discourse in acquired and progressive neurological disorders can improve specificity and modification of treatment and goal-setting, for example, identification of the extent to which treatments generalise to discourse and mechanisms of generalisation (Boyle, 2011, 2020).

Barriers to analysing aphasic spoken discourse

While discourse analysis has many benefits, considerable implementation barriers have been identified in research and clinical settings (Bryant et al., 2017; Cruice et al., 2020; Stark, Dutta, Murray, Fromm et al., 2021). The most common barriers include (1) outdated assessments which do not take into account modern social norms and lack representation, such as the Cookie Theft Picture Description (although for an update, see Berube et al., 2019); (2) time constraints; and (3) lack of training and tools (Bryant et al., 2017; Cruice et al., 2020; Stark, Dutta, Murray, Fromm et al., 2021). These barriers have likely resulted in the limited implementation of discourse analysis (Bryant et al., 2017; Cruice et al., 2020; Stark, Dutta, Murray, Fromm et al., 2021). Outdated assessments may lead to decreased tellability (Olness & Ulatowska, 2011), and at their worst perpetuate long-standing stereotypes and biases (e.g., Berube et al., 2019). The amount of time required to transcribe (some estimate 10–15 min to transcribe per 1 min of aphasic speech; Boles, 1998), analyse and interpret data is a barrier in both clinical and research settings, though clinical practitioners may feel that this barrier is larger (Stark, Dutta, Murray, Fromm et al., 2021).

Indeed, in a survey of 162 clinicians and researchers, 93.8% of respondents cited time as a barrier for discourse analysis, a finding that dwarfed the other barriers cited in the study (skills and knowledge, training, access to tools and resources, confidence, protocol application, results interpretation) (Stark, Dutta, Murray, Fromm et al., 2021). This finding was echoed by two other surveys, which also reported time as the biggest barrier to discourse analysis (41.5% in an international speech-language pathologists [SLP] sample, Bryant et al., 2017; 78% in a UK SLP sample, Cruice et al., 2020).

Another critical issue in discourse analysis is related to the complexity of collecting, analysing and then interpreting the results. Researchers and clinicians cite a lack of training in discourse elicitation and analysis in formal coursework or professional settings (Bryant et al., 2017; Cruice et al., 2020; Stark, Dutta, Murray, Fromm et al., 2021), which likely leads to fewer instances of discourse analysis. Bryant et al. (2017) report that only 30% of respondents (speech therapists) agreed or strongly agreed that they felt competent using discourse analysis to assess language in aphasia, despite 50% agreeing or strongly agreeing that detailed linguistic analysis of discourse is important for the assessment of language in aphasia. The feeling of reduced competency may be related to the complexity of discourse (i.e., involving many language and nonlinguistic behaviours), as well as the complex relationship between the levels of discourse, as highlighted previously. Clinicians, more often than researchers, cite inadequate training and access to tools and resources as barriers to discourse collection, analysis, and interpretation (Stark, Dutta, Murray, Fromm et al., 2021). This suggests that the complexity of discourse extends from collection through to interpretation, such that clinicians in particular may not receive the training or continuing education required to feel confident and competent in employing discourse analysis in their practice with individuals with aphasia.

Despite Bryant et al. (2016) 's systematic review identifying more than 500 different metrics of discourse in studies, such as number of tokens, number of utterances and noun/verb ratio, both clinicians and researchers cite limited tools for discourse analysis (Stark, Dutta, Murray, Fromm et al., 2021). This is likely because most of Bryant et al.'s (2016) evidence for the >500 discourse metrics were extracted from transcriptions. As such, the large number of metrics likely reflects a predisposition to analyse transcripts using programs or in-house methods, rather than a set tool. Further, survey findings suggest that existing tools may demonstrate limited evidence for psychometric soundness in validity, reliability and feasibility (Pritchard et al., 2018; Stark, Dutta, Murray, Fromm et al., 2021). Indeed, in the Stark, Dutta, Murray, Fromm et al. (2021) survey of researchers and clinicians, a majority of respondents (81.8%) felt there were inadequate psychometric data for spoken discourse measures and tools. Respondents cited that there were major barriers to collecting psychometric data, such as time (82.7%), knowledge and training (60.9%), funds (46.4%) and personnel (42.7%) (Stark, Dutta, Murray, Fromm et al., 2021). Despite these issues, respondents described psychometric properties of discourse data as important for comparing and interpreting discourse measures across individuals and approaches (e.g., 'Without adequate psychometric properties described, interpretation of results is problematic, and clinical application of measures will be limited') and that psychometric data acquisition was best practice (Stark, Dutta, Murray, Fromm et al., 2021).

Establishing psychometric properties of tools are necessary to instil confidence in users (e.g., clinicians, researchers) that the task and outcomes are reliable and valid for meaningful decision making. Reliability includes a variety of components, including test-retest reliability (the extent to which the measure, and its outcomes, produces a similar result at two different time-points) and rater reliability (the extent to which raters, or the same rater, use the tool similarly across two different administrations). In Classical Test Theory, a lack of reliability means a higher rate of error, which leads to a lack of confidence in the tool. In the worst cases, this yields a tool that cannot sensitively measure a behaviour or a change in behaviour across time (Stark et al., 2023). Validity, which is closely related to reliability, pertains to the well-foundedness of the tool, such that it measures a logical construct (construct validity), its outcomes are logically related to other similar outcomes (concurrent validity) and not to dissimilar outcomes (discriminant or divergent validity), and that it adequately differentiates groups of individuals known to differ on specific outcomes (known groups validity) (Zumbo, 2009). Tools must be both valid and reliable to have high clinical utility.

Finally, there is limited implementation of tools developed in research settings into clinical practice. This may be because most tools are developed in and for research contexts which may not translate well into clinical environments, and/or which may not measure outcomes that clinicians want or need to measure. For example, there is a proclivity in transcription-oriented research to focus on linguistic outcomes, whereas many clinicians also want to understand more functional communication outcomes extending beyond linguistic features (Stark, Dutta, Murray, Fromm et al., 2021). An example of a functional outcome might be the ability of the person to maintain topic during a conversation, which is at the level of macrostructural/planning or pragmatic levels. In the Cruice et al. (2020), survey of practising SLPs in the United Kingdom, 75% of respondents suggested that they would be inclined to implement a discourse assessment-to-goal-setting process (i.e., including discourse analysis and interpretation) if it took ≤60 min. This finding suggests that limiting factors to implementation include time efficiency as well as time-saving tools and adequate training.

Transcription versus transcription-less

The gold standard for analysing spoken discourse is to transcribe the speech and then conduct analyses on it. This is because transcription—be it orthographic or phonetic—enables quantification of discourse behaviours at a fine-grained level. More broadly, transcription can help researchers and clinicians understand their data better and facilitate further research and use of the data. The primary drawback of transcription is its time-intensive nature: it takes time to transcribe speech because automatic speech recognition is not yet able to handle aphasic speech without considerable manual editing. Even after the initial transcription is completed, additional time is needed to check the transcription, add manual codes (e.g., error types) and interpret the data.

Transcription-less methods are thought to overcome some time limitations by providing a tool that can be completed online, that is, whilst in the presence of the client/participant and without transcription. Transcription-less tools and analysis procedures have been referred to as judgement-based analysis, perceptual analysis, subjective analysis and without transcription analysis across studies. Transcription-less tools are desirable because they have the potential to address the barriers of time and tool availability. However, if the tools do not measure what constituents want to measure (i.e., SLPs may want to capture specific aspects of discourse that will enable them to more thoroughly identify functional goals), and are not reliable or valid, then they will not fully address the barriers discussed.

Transcription-less methods depend upon in-the-moment perceptions of behaviour. However, if a recording is collected, perceptual tools can be used to score the discourse offline. Several standard aphasia batteries include a perceptual rating scale for connected speech analysis: Boston Diagnostic Aphasia Examination (BDAE; Goodglass & Kaplan, 1972), Western Aphasia Battery—Revised (WAB-R; Kertesz, 2007), and Quick Aphasia Battery (Wilson et al., 2018). These batteries also provide space for the assessor to transcribe the speech, but it is not explicitly stated that transcription is needed to use the rating scales. As described by several international surveys, clinicians and researchers typically use perceptual judgement when employing these scales, and these scales are also the most commonly used transcription-less method for discourse analysis (Bryant et al., 2017; Stark, Dutta, Murray, Fromm et al., 2021).

An issue with existing rating scales in comprehensive, standardised aphasia batteries is that they focus mostly on linguistic and propositional levels (e.g., fluency, speech rate, semantic content), and do not evaluate macrostructural/planning or pragmatic levels. For example, in the Comprehensive Aphasia Test (Swinburn et al., 2004), the spoken picture description component is rated by evaluating appropriate and inappropriate word counts (including errors), syntactic variety, grammatical well-formedness and speed ratings. These are predominantly linguistic parameters. Because these standard batteries collect limited samples and genres (typically, a very short interview such as ‘Have you been here before?’ and a short picture description; e.g., Kertesz, 2007), it is difficult to assess macrostructural/planning and pragmatic levels, and therefore it is not surprising that the rating scales do not evaluate those components of the discourse.

There have also been some issues reported with reliability of the rating scales on standard batteries. Trupe (1984) noted that the inter-rater reliability of scoring the picture description on the WAB was relatively low. Specifically, she criticised the rating of the Fluency score, which is critical for assigning aphasia type in this battery. A variation of one point in the Fluency score can result in a different diagnostic classification, between a fluent (score of five or more) and non-fluent (score of four or less) subtype. This is likely because fluency is a concept that encompasses many things, such as motor speech programming and lexical-semantic and phonological access. It is therefore interpreted differently by both clinicians and researchers and is likely influenced by clinical experience (Clough & Gordon, 2020; Gordon & Clough, 2020). Further, as Trupe (1984) rightly points out, having scale-type variables that are all encompassing does not allow for more fine-grained analysis. On the WAB-R, a Fluency rating of four is described as, ‘Halting, telegraphic speech; mostly single words; paraphasias; occasional prepositional phrases; severe word-finding difficulty; no more than two complete sentences with the exception of automatic sentences; characteristic of agrammatic, non-fluent aphasia’ while a Fluency rating of five is described as, ‘often telegraphic but more fluent speech with some grammatical organisation; marked word-finding difficulty; paraphasias may be prominent; few, but more than two propositional sentences’. When a person whose discourse is largely telegraphic is also able to produce islands of fluent, jargon-like speech, the rating scale points become blurred and less easily differentiable. One could also argue that, in those cases, the rating scale has limited validity, in that its diagnostic utility is limited. Recent work suggests that categorical ratings (fluent versus non-fluent, rather than the WAB-R's Likert-scale rating) may have better inter-rater agreement (Metu et al., 2023). Given these factors, rating scales on standardised aphasia batteries are likely not sufficient for a comprehensively informative discourse analysis, especially since they typically disregard the behaviours unique to discourse observed at the macrostructural/planning and pragmatic levels.

Rationale for this scoping review

The impetus for this scoping review is: (1) the limited clinical implementation of discourse analysis (Bryant et al., 2017; Cruice et al., 2020; Stark, Dutta, Murray, Fromm et al., 2021), despite research efforts to reduce barriers such as time and training; and (2) that when practising SLPs *do* collect discourse samples, they report using formal and informal perceptual judgements of discourse without transcription, rather than transcribing and using quantitative measures presented throughout the literature (Bryant et al., 2017; Cruice et al., 2020; Stark, Dutta, Murray, Fromm et al., 2021). This scoping review's rationale was, therefore, to identify available evidence for transcription-less tool reliability and validity specific to aphasic discourse and beyond those included on standardised batteries. The intent was to characterise in greater detail the procedures for transcription-less tool use and validation (e.g., rater and participant characteristics; types of features included on the tool), and to identify and analyse knowledge gaps related to transcription-less tool use for research and for use clinically.

Objectives and focus of the review

The primary question posed by this scoping review is: 'What transcription-less tools and analysis procedures are available to assess discourse in people with aphasia?' Five sub-questions elaborate on specific components of interest:

1. What types of measurements (e.g., categorical, Likert-style) do transcription-less tools employ?
2. What information (as categorised by the four Dipper et al., 2021, levels, and the additional inclusion of a non-verbal/paralinguistic level and a non-categorizable level) do transcription-less tools measure?
3. What are the characteristics of speakers with aphasia included in these studies?
4. What are the characteristics of the raters included in these studies?
5. What psychometric properties (reliability, validity) are available for the tools, and have the tools been investigated for their potential to be implemented in a clinical setting?

The outcome of this scoping review is to summarise the current state of the evidence and discuss next steps for improvement of transcription-less analysis for aphasic discourse. The purpose of this scoping review is not to offer an exhaustive or systematic review of all tools available for discourse analysis.

METHOD

This scoping review adhered to the guidelines developed by the Joanna Briggs Institute Scoping Review Methodology Group (Peters et al., 2020). Scoping reviews are defined as 'exploratory projects that systematically map the literature available on a topic, identifying key concepts, theories, sources of evidence and gaps in the research' (Grimshaw, 2010). Since there has been limited synthesis of transcription-less discourse analysis research, a scoping review to map the evidence was deemed more appropriate than a systematic review that would address a constrained empirical question. While this scoping review was not preregistered, it followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews checklist to ensure rigour in reporting (see Supplementary materials 1; note: pre-registration is not a requirement of a scoping review, per Peters et al., 2020).

Scoping the literature and searching processes

This scoping review's key population of interest is individuals with acquired, non-progressive aphasia. Therefore, articles studying other populations were excluded. Further, included articles needed to focus on discourse measurement. In order to be maximally inclusive of the research, Armstrong's (2000) definition of discourse was used: elicited language above a single simple clause used for a specific purpose. Defining what is meant by transcription-less or perceptual discourse analysis proved more challenging than defining the population of interest and what is meant by discourse. There has been very little uniformity in the terminology used to describe the analyses of interest in this review. Therefore, perceptual discourse analysis was conceptualised as any analysis which primarily relied on listener impressions and lacked numerical quantification of specific discourse behaviours (e.g., counts, frequency, percentages). Transcription-less analyses were defined as any discourse analysis which did not require a written record of the discourse sample in order to be completed. This definition enabled inclusion of analyses which relied upon audio- or video-recordings of the discourse sample. Inclusion was not limited based on the research context beyond requiring that research articles were written in English in order to be reviewed by the research team. Therefore, evidence was reviewed across cultural, geographic and linguistic contexts, as long as they aligned with the concepts outlined here. Finally, there was no restriction on publication date for inclusion.

Search strategy

The scoping review was conducted between the months of October 2022 and January 2023 on PubMed/NCBI, Academic Search Complete and Linguistics and Language Behavior Abstracts. The following exact search terms (in 'text word') were employed to search in each of theseatabases:

- Discourse OR 'connected speech' OR 'spontaneous speech' OR narrat* OR storytelling OR story-telling OR 'story telling' OR 'picture description' OR 'picture exposition'

  AND

- Aphasi* OR dysphasi*

  AND

- Perceptual OR scoring OR subjective OR rubric OR 'rating scale' OR 'no transcription' OR transcription-less OR transcription-less OR 'without transcription' OR 'judgment-based' OR 'judgment based' OR 'listener perception'.

The terms 'conversation' and 'dialogue' were not explicitly included since these behaviours fall under the broad category of discourse and many papers were found involving both dialogue and monologue with thesesearch terms. Reference lists of relevant papers were evaluated to ensure that the search was comprehensive.

Selection and appraisal of documents

Peer-reviewed, published, stand-alone tools that have been devised specifically to evaluate discourse elicited from an individual with aphasia were the subject of this scoping review. For this reason, articles were excluded if they met the following parameters:

- Papers published in languages other than English since the authors are monolingual English speakers.
- Rating scales or tools from standardised batteries, since these scales often have limited granularity, generally do not consider language behaviours unique to discourse and are not created to stand alone as tools.
- Tools specialised to evaluate progressive aphasia (e.g., primary progressive aphasia) or cognitive-communicative function after other brain injury (e.g., traumatic brain injury).
- Conference papers or conference proceedings and unpublished theses or dissertations. These scholarly outputs were excluded since they typically lack full peer review, and in the case of conference papers/proceedings, have limited detail and specificity.

After initial screening and prior to the data extraction process, articles were further excluded if they had the following characteristics:

- Any analyses that required specialised software such as automatic speech recognition (e.g., Croteau et al., 2018; Dalton et al., 2022; Qin et al., 2020). These were excluded since they either would not be widely accessible by clinicians, required transcription or were very time intensive.
- Studies with insufficient methodological details, such that they would not be replicable. Goodman et al. (2016) proposed a new lexicon for research reproducibility, highlighting *methods reproducibility* as providing sufficient detail about procedures and data so that the same procedures could be exactly repeated. For example, one excluded study, Armstrong et al. (2007), wrote that a transcription-less tool was used by students to rate several measures (e.g., gesture use), across subcategories (e.g., ideographs, deictic movements), and across several tasks (e.g., picture description, narrative). Original authors gave no details about the tool (e.g., Likert scale, visual analogue, etc), such that rater reliability results, and the analyses, could not be thoroughly examined and the procedures not replicated. In another excluded example, Ulatowska et al. (2001), it was unclear if the discussed measures came from a transcription-less analysis or a transcription analysis, thus dampening confidence that a transcription-less tool was used.
- Articles describing batteries evaluating general functional communication (e.g., Communicative Effectiveness Index, Lomas et al., 1989). These were excluded as they do not directly evaluate discourse and are more holistic in nature.

Analysis processes conducted on each included paper and specific analysis objectives

When papers met inclusion criteria, further analyses were conducted on the tools being described in the papers, guided by the five scoping review objectives. Authors B.C.S. and S.G.D. independently identified evidence for each objective in the full papers that were reviewed, and where disagreements existed, reached consensus via discussion. In addition to extracting data to evaluate these objectives, the following data were extracted from every included article and can be found in Supplementary material 2: full citation, country of study, context of study (university, clinic), aim/purpose of study, summary of study outcome and a description of the rating tool (including details about the rating scale, length of time to use the tool and ease of using the tool).

Objective 1: Types of transcription-less tool analysis procedures

The first goal was to categorise the analysis tools into categories to enable more streamlined understanding of tool mechanisms. Authors discussed and came to the consensus that there were two primary types of tools: those using categorical rating scales and those using direct magnitude estimations. Categorical scales rate or identify features based on categories, for example, presence/absence, accurate/inaccurate or Likert-type scales. Direct magnitude estimations (DME) give numerical estimates to a single stimulus, and then subsequent stimuli are rated in relation to the first stimuli (e.g., Doyle et al., 1996). As such, identified studies were organised based on whether the tools fit into either categorical or DME types.

While reviewing the articles, patterns were identified that enabled further categorization of tools: (1) the same group of authors, or similar groups of authors, published similar rating scales across several papers, often in different contexts (e.g., procedural analysis versus narrative analysis); (2) several author groups published similar styles of analyses; and (3) in rarer cases, articles contained two different types of tools, that is, one of DME type and one of categorical. In the first case, all similar author groups and similar tool types were grouped under a single category (DME or categorical), as appropriate. In the second case, similar tools, even if reported by different author groups, were grouped under a single category as appropriate. In the third case, tools were grouped into the most appropriate category. This streamlined identification of available transcription-less analysis procedures. If there was a published name for the tool, that name was used when describing it. If there was no official name to the analysis, a label was created for it based on its description in the article in order to simplify the presentation of results.

Objective 2: What type of information was extracted from the discourse

For this objective, data extraction included: the discourse level(s) evaluated (i.e., linguistic, propositional, macrostructural/planning, pragmatic, paralinguistic/non-verbal and non-categorizable), the type (monologue, dialogue) and the genre of discourse for which the analysis was developed. Importantly, each discourse level interacts with every other level, and oftentimes a concept can be measured across levels. For example, informativeness could be assigned to the linguistic level (e.g., correct information units, core lexicon analysis), the propositional level (e.g., main concept analysis, main event analysis), or the pragmatic level (e.g., judgements about the overall informativeness of a discourse sample). Therefore, it was sometimes necessary to make decisions regarding the level of discourse with which an item or scale was best aligned. In these instances, B.C.S. and S.G.D. relied upon the definitions provided by the articles' authors and the specific content of the rating item/scale to make a determination.

The Supplementary material delineates how each article's tool(s) aligned with the Dipper et al. (2021) framework and notes whether tool(s) evaluated any non-verbal or paralinguistic features, or if any features on the tool were otherwise not categorizable. Not categorizable features tended to be those broadly evaluating aspects of speech and fluency (e.g., pauses; hesitations) that did not directly map to the theoretical framework, or to non-verbal/paralinguistic behaviours. Supplementary Material also includes the type of discourse to which the tool was applied (dialogue, monologue), its potential to be used for other types of discourse, the genre of discourse to which the tool was applied and its potential to be used for other genres of discourse.

Objective 3: Characteristics of speakers with aphasia across studies

Data acquisition to satisfy this objective included demographic data (age, sex, race/ethnicity, aetiology, languages spoken and any other available information), as well as the aphasia type and severity of individuals with aphasia on which the transcription-less tools were used. Since there are some disagreements about how to classify aphasia types, this scoping review will report the types and severities reported by the individual articles.

Objective 4: Characteristics of raters/tool users

Three categories of raters were defined: naive, learner and expert. Naive included raters with no formal education or exposure to individuals with aphasia. Learners included students in undergraduate, master's and doctoral programs in speech-language pathology/communication sciences and disorders, with or without specific experience working with individuals with aphasia. Experts included certified SLPs, other professionals, or researchers with expertise working with individuals with aphasia. Studies classified these three categories slightly differently; therefore, verbiage from the study itself was used to describe the raters whenever possible. In addition to rater classification, the Supplementary Material reports on the specifics of the raters, including demographics, training for using the tool, general tool use procedures for raters and the method of data capture/analysis (i.e., audio/visual).

Objective 5: Psychometric properties and implementation

Assessing the quality of the reported psychometric data generally falls outside the purview of a scoping review, thus the psychometric properties are reported as the original authors interpreted them (e.g., 'strong' inter-rater reliability). The extent to which the tool was or could be used in a clinical setting (implementation) was also evaluated. This included whether authors had asked any questions to raters regarding the tool's potential to be implemented in a clinical setting. In the Supplementary Material, the statistics reported by each paper about reliability and validity are provided, and descriptive information is provided about implementation.

Summary of data extraction

In order to achieve our five objectives, the following data were extracted for each tool:

- Country and context of study (university, clinic).
- Aim/purpose of study and summary of study outcome.
- Description of the tool (including specific questions if provided, the rating scale, length of time it took to use the tool, ease of using the tool).
- Type of tool (categorical versus DME).
- Levels of discourse evaluated by each tool: linguistic, propositional, macrostructural/planning, pragmatic, non-verbal or paralinguistic and not categorizable.
- Discourse type (monologue versus dialogue) the tool has been and could be used for.
- Discourse genre (e.g., narrative, interview) the tool has been and could be used for.
- Descriptive information about the sample of individuals with aphasia from whom the discourse was elicited.
- Descriptive information about the raters/users of the tool (e.g., naivety, number), and how information was presented to raters (e.g., audio, visual).
- Training and procedures for using the tool.
- Psychometric properties of the tool, including reliability, validity and implementation.

RESULTS

A total of 396 articles were identified: 66 articles from PubMed/NCBI, 289 from Academic Search Complete, and 41 from Linguistics and Language Behaviour Abstracts (Figure 1). After B.C.S. identified 28 full articles from the databases to review based on their abstracts, a further 11 articles were identified by B.C.S. and S.G.D. from reference lists. Therefore, 35 articles were fully read/inspected independently by both S.G.D. and B.C.S. for inclusion/exclusion criteria. Ultimately, 21 articles met inclusion/exclusion criteria. See Supplementary material 3 for 18 excluded, fully reviewed articles and the reasons for exclusion.

Objective 1: Types of transcription-less tool analysis procedures

Eleven distinct types of categorical scale analyses (16 total papers) and three distinct types of DME scale analyses (five total papers) fit inclusion/exclusion criteria (Tables 1 and 2).

The categorical scales/analyses were core lexicon analysis (CoreLex; Dalton et al., 2020; Kim & Wright, 2020), Story Retell Procedure-derived information unit (IU; Hula et al., 2003), morphosyntactic analysis (Ballard & Thompson, 1999), main concept analysis (MCA; Dalton et al., 2020), Auditory-Perceptual Rating of Connected Speech in Aphasia (APROCSA; Casilio et al., 2019), listener perception rating scales (Behrns et al., 2009; Cupit et al., 2010; Harmon et al., 2016; Ross & Wertz, 1999), Discourse Abilities Profile (DAP; Terrell & Ripich, 1989), speech function rating scale (Copeland, 1989), conversation communication strategies (Herrmann, 1989), morphosyntactic analy-

TABLE 1 Perceptual/transcription-less scales with categorical outcomes.

Author(s) and year	Analysis	Description of rating scale	Discourse genres	Discourse level	Data capture
Dalton et al. (2020)	Core Lexicon Analysis (CoreLex)[a]	*Summary*: Listeners used a checklist to score the presence or absence of CoreLex from AphasiaBank tasks. *Method of Rating*: Categorical (2) *Speakers*: 15 participants with chronic aphasia with a range of types and severities; latent included *Raters*: Research team members	Monologue—task specific across multiple genres	Linguistic	Video recording
Kim and Wright (2020)	Core Lexicon Analysis (CoreLex)[a]	*Summary*: Listeners used a checklist to score the presence/absence of CoreLex from wordless picture books. *Method of Rating*: Categorical (2) *Speakers*: 11 participants with chronic aphasia with a range of types and severities; latent excluded *Raters*: 4 experienced students	Monologue—task specific across multiple genres	Linguistic	Audio recording
Hula et al. (2003)	Story Retell Information Units (IU)*	*Summary*: Using Story Retell Procedure (McNeil et al., 2001), judges used a pre-developed information unit checklist to score information units during narratives. *Method of Rating*: Categorical (2) *Speakers*: 4 participants with aphasia, 11 healthy controls; inclusion of latent unknown *Raters*: 4 raters (2 SLPs & PhD students, 1 psychologist, 1 MA student)	Monologue—task specific across multiple genres	Linguistic	Audio recording
Dalton et al. (2020)	Main Concept Analysis (MCA)[a],[b]	*Summary*: Listeners used a checklist to score main concepts from AphasiaBank tasks for accuracy and completeness. *Method of Rating*: Categorical (5) *Speakers*: 15 participants with chronic aphasia with a range of types and severities; latent included *Raters*: Research team members	Monologue—task specific across multiple genres	Propositional	Video recording
Casilio et al. (2019)	Auditory—Perceptual Rating of Connected Speech in Aphasia	*Summary*: Listeners rated 27 language features for use with monologic discourse tasks. *Method of Rating*: 5-point Likert *Speakers*: 25 participants with chronic aphasia with a range of types and severities; latent excluded *Raters*: 3 experienced, 12 second-year MA students	Monologue	Linguistic Propositional Macrostructural Pragmatic Uncategorizable	Video recording
Terrell and Ripich (1989)	Discourse Abilities Profile (DAP)	*Summary*: This transcription-less scoring was devised by authors, but no data was presented by authors regarding its use in context. The DAP includes presence/absence of a variety of behaviours, separated by monologue genres (narrative, procedural) and for dialogue. *Method of Rating*: Categorical (2); 5-point Likert *Speakers*: N/A *Raters*: N/A	Monologue and dialogue	Propositional Macrostructural Pragmatic Non-verbal	N/A

(Continues)

TABLE 1 (Continued)

Author(s) and year	Analysis	Description of rating scale	Discourse genres	Discourse level	Data capture
Harmon et al. (2016)	Listener rating/perception	*Summary*: Listeners rated narratives on 9-items probing perceptions about speech output, speaker attributes and listener feelings. *Method of Rating*: 7-point Likert *Speakers*: 6 participants with Broca's aphasia; 3 healthy controls *Raters*: 18 undergraduate students; 18 graduate students	Monologue and dialogue	Pragmatic, Uncategorizable	Audio recording
Ross and Wertz (1999)	Listener rating/perception	*Summary*: Listeners compared two time-points of picture scene descriptions and judged whether the second sample they heard was better than, the same as, or worse than the previous in terms of communicative ability. *Method of Rating*: Categorical (3) *Speakers*: 22 participants with sub-acute to chronic aphasia with a range of severities and types; latent excluded *Raters*: 10 graduate students	Monologue and dialogue	Uncategorizable	Video recording
Behrns et al. (2009)	Listener rating/perception	*Summary*: Listeners rated general impressions of narrative (e.g., difficult/easy to understand), vocabulary (e.g., inadequate/adequate), structure (e.g., incoherent/coherent), and overall impression of the narrator (e.g., seemed to dislike/like telling the story). *Method of Rating*: 100 mm VAS *Speakers*: 8 Swedish speakers with chronic aphasia with a range of types and severities; latent excluded *Raters*: 60 listeners with unknown experience	Monologue and dialogue	Linguistic, Macrostructural, Pragmatic	Video recording
Cupit et al. (2010)	Listener rating/perception	*Summary*: Listeners rated pre- and post-therapy Cinderella narratives for four discourse parameters (amount of information, ability to transmit the message, ability to find the words, and degree of ease in retelling the narrative). *Method of Rating*: 7-point Likert *Speakers*: 11 participants with chronic aphasia with a range of types and severities; latent excluded *Raters*: 10 SLPs; 10 naive younger adults; 10 naive older adults	Monologue and dialogue	Linguistic, Propositional, Uncategorizable	Audio recording
Copeland (1989)	Speech function rating scale	*Summary*: Listeners rated 20 speech functions (e.g., greetings, informing, explaining, arguing, requesting) from conversation. *Method of Rating*: Ordinal (6) *Speakers*: 10 individuals with chronic Broca's aphasia *Raters*: 3 SLPs	Dialogue	Pragmatic Non-verbal Uncategorizable	Audio recording
Herrmann (1989)	Conversation communication strategies	*Summary*: Listeners rated communication impairment and communication strategy use during conversation. *Method of Rating*: Ordinal (5 and 7 levels) *Speakers*: 20 German speakers with moderate-severe non-fluent aphasia *Raters*: Study team	Dialogue (Communication Impairment Scoring may work for monologue)	Propositional Pragmatic Non-verbal	Unclear, but may have been online

(Continues)

TABLE 1 (Continued)

Author(s) and year	Analysis	Description of rating scale	Discourse genres	Discourse level	Data capture
Ballard and Thompson (1999)	Morphosyntactic analysis	*Summary*: Listeners rated narratives based on (a) content, (b) coherence, (c) fluency and efficiency of expression, (d) length and complexity of sentences used, and (e) grammaticality of utterances. *Method of Rating*: Numerical scale, unclear if Likert *Speakers*: 5 speakers with chronic moderate Broca's aphasia *Raters*: 10 speech-language pathology graduate students	Monologue and dialogue	Linguistic Propositional Macrostructural Uncategorizable	Audio recording
Ulatowska 1981;	Subjective ratings of content, coherence, and clarity	*Summary*: Listeners rated the content and clarity of procedures and narratives. *Method of Rating*: Categorical (2 or 3) *Speakers*: 10 speakers with chronic aphasia of differing severities and types; 10 healthy control speakers *Raters*: 5 doctoral speech-language pathology students unfamiliar with the speakers	Monologue, largely narrative and procedural genres	Propositional, Macrostructural	Audio recording
Ulatowska et al. (1983a & b)	Subjective ratings of content, coherence, and clarity	*Summary*: Listeners rated the content and clarity of procedural and narrative tasks. *Method of Rating*: Categorical (3 or 4) *Speakers*: 15 speakers with sub-acute to chronic moderate aphasia of various types; 15 healthy control speakers *Raters*: 3 SLPs unfamiliar with the speakers	Monologue, largely narrative and procedural genres	Propositional, Macrostructural	Audio recording
Ulatowska et al. (2013)	Subjective ratings of content, coherence, and clarity	*Summary*: Listeners rated the level of coherence and clarity of narratives. *Method of Rating*: Categorical (3) *Speakers*: 16 speakers with sub-acute to chronic mild or moderate aphasia of various types, latent excluded *Raters*: 5 raters, no other details provided	Monologue, largely narrative and procedural genres	Propositional, Macrostructural	Audio recording
Ulatowska et al. (2003)	Subjective rating of discourse quality	*Summary*: Listeners rated the quality of narratives along three dimensions: coherence, reference, and emplotment. *Method of Rating*: Categorical (5) *Speakers*: 12 African American speakers with mild or moderate aphasia of various types, latent excluded *Raters*: 6 raters, no other details provided	Monologue	Macrostructural	Unclear

[a] The primary analysis is transcription based, but a secondary analysis was completed using audio and/or video recordings only.

[b] While many other groups have published on this measure, this was the only study that reported a transcription-less scoring approach.

Abbreviations: MA, master of arts; SLP, speech-language pathologist; VAS, Visual Analogue Scale.

TABLE 2 Perceptual/transcription-less scales with direct magnitude estimations.

Author(s) and Year	Analysis	Description of rating scale	Discourse genres	Discourse level	Data capture
Hickey and Rondeau (2005)	Listener perception of conversation	*Summary*: Raters assessed seven dimensions of conversation (e.g., overall quality of conversation, topics of conversations continued until both persons appeared ready to change it, equal turn-taking) to examine the clinical significance of training-related changes in conversations between a student volunteer and an elder with aphasia. *Method of Rating*: Direct magnitude estimation *Speakers*: 1 speaker with chronic aphasia *Raters*: 10 naive raters; 10 second-year speech-language pathology graduate students; 10 speech-language pathologists with at least 3 years of experience with aphasia	Dialogue	Propositional Macrostructural Pragmatic	Video recording
Doyle et al. (1996)	Informativeness	*Summary*: Listeners rated overall informativeness in one sample (anchor) and then compared subsequent samples to the anchor (i.e., more or less informative) in picture scene/picture sequence descriptions and procedural tasks. *Method of Rating*: Direct magnitude estimation *Speakers*: 25 speakers with aphasia of various types and severity, latent included with anomic *Raters*: 11 naive older raters	Monologue	Propositional	Audio recording
Jacobs (2001)	Informativeness	*Summary*: Listeners provided ratings (with an anchor system as described above) for 4 narrative constructs: effectiveness, informativeness, grammaticality, and listener comfort. *Method of Rating*: Direct magnitude estimation *Speakers*: 5 speakers with chronic mild to moderate Broca's aphasia *Raters*: 10 speech-language pathology graduate students unfamiliar with the speakers	Monologue	Linguistic, Propositional, Macrostructural, Pragmatic	Audio recording
Webster and Morris (2019)	Informativeness	*Summary*: Listeners rated the anchor for informativeness, and then rated each successive sample compared to the anchor during picture scene descriptions. *Method of Rating*: Direct magnitude estimation *Speakers*: 20 speakers with sub-acute to chronic aphasia of varying types and severities *Raters*: 11 university students with little to no aphasia experience	Monologue	Propositional	Audio recording
Copeland (1989)	Burden, speed, and paralinguistic features of conversation	*Summary*: Speech-language pathologists rated three conversations for (1) burden of conversation; (2) speed and success; and (3) paralinguistic features (amount and helpfulness; two scales). *Method of Rating*: Direct magnitude estimation *Speakers*: 10 individuals with chronic Broca's aphasia *Raters*: 3 speech-language pathologists	Dialogue	Pragmatic Non-verbal Uncategorizable	Audio recording

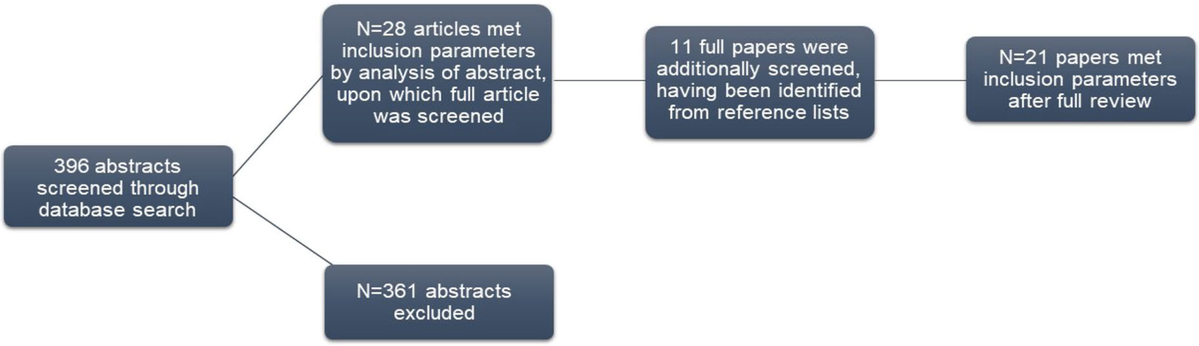

N=28 articles met inclusion parameters by analysis of abstract, upon which full article was screened
11 full papers were additionally screened, having been identified from reference lists
N=21 papers met inclusion parameters after full review
396 abstracts screened through database search
N=361 abstracts excluded

FIGURE 1 Consolidated Standards of Reporting Trials diagram of scoping review. [Colour figure can be viewed at wileyonlinelibrary.com]

sis (Ballard & Thompson, 1999), ratings of content, clarity and coherence (Ulatowska et al., 1981, 2013; Ulatowska, Doyel et al., 1983; Ulatowska, Freedman-Stern et al., 1983), and ratings of quality (Ulatowska et al., 2003). For categorical analyses, the majority used Likert-type scales. For example, nearly all of Ulatowska and colleagues' combined works used numerical scales, ranging from three to five points. A few studies used dichotomous categorical choices (e.g., presence/absence for core lexicon analysis [Dalton et al., 2020; Kim & Wright, 2020], DAP [Terrell & Ripich, 1989], and information unit analysis [e.g., Doyle et al., 1996]; bad/good, interesting/not interesting, difficult/easy to understand from Behrns et al., 2009). MCA (Dalton & Richardson, 2019) used dichotomous decisions about accuracy and completeness to assign utterances to one of five categories (ranging from accurate/complete to absent).

The three types of DME scale analyses were listener perceptual ratings of conversation (Hickey & Rondeau, 2005); informativeness (Doyle et al., 1996; Jacobs, 2001; Webster & Morris, 2019); and burden, speed and paralinguistic features of conversation (Copeland, 1989).

Objective 2: Main discourse components being evaluated

Each transcription-less analysis was categorised by the aspects of discourse that it was evaluating, using the levels identified by Dipper et al. (2021), with the additional options of non-verbal/paralinguistic and non-categorizable. Most transcription-less analyses evaluated discourse across levels rather than focusing on a single level. Figure 2 provides a visualisation of the discourse components evaluated by each study. See Supplementary Materials for a more detailed description of the features employed by each study, and into which discourse level they fit. Approximately 38% of features used across all tools were pragmatic, ~19% were linguistic, ~16% were propositional, ~12% were macrostructural/planning, ~4% were non-verbal/paralinguistic and ~10% were not categorizable.

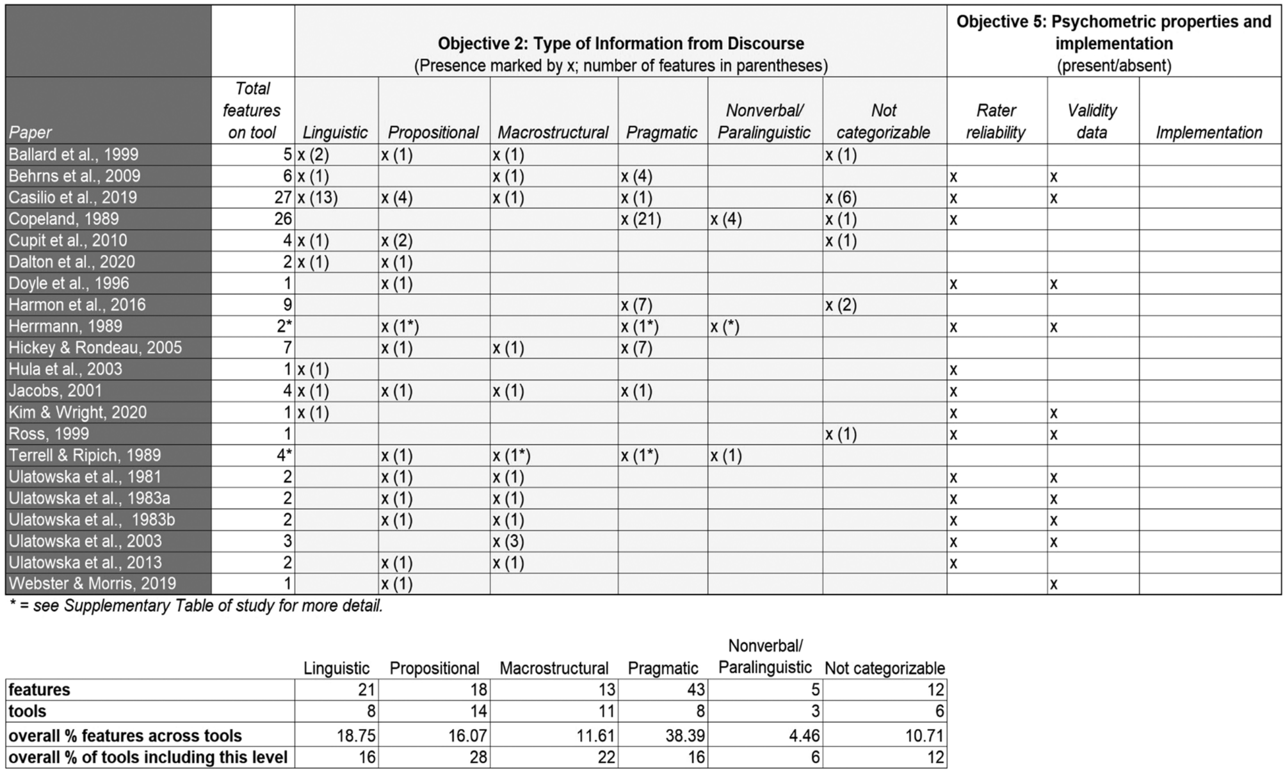
Linguistic

Of the 21 papers evaluated, eight included tools containing at least one linguistic feature (38%). examples of linguistic features included in the papers follow.

In core lexicon analysis (Dalton et al., 2020; Kim & Wright, 2020), the presence/absence of stimulus-relevant lemmas (as derived from a normative sample) evaluated typicality of vocabulary usage and may provide some information about lexical-semantic access and syntax. Hula et al. (2003) evaluated lexical-semantic informativeness by asking judges to rate presence/absence of predefined information units for the Story Retell Procedure. Two other studies evaluated lexical-semantic ability by asking listeners to rate a patient's ability to 'find adequate words' (Behrns et al., 2009; Cupit et al., 2010) and one study had listeners provide ratings of grammaticality (e.g., syntax; Jacobs, 2001). One study evaluated components that they broadly described as being morphosyntactic (Ballard & Thompson, 1999), having judges perceptually rate narrative samples from a larger treatment study on numerical scales for length and complexity of sentences used and grammaticality of utterances. Finally, the APROCSA (Casilio et al., 2019) had listeners rate 27 language behaviours including 13 linguistic features such as presence and estimated frequency of semantic and phonemic paraphasias, and presence of short and simplified utterances.

Propositional

Of the 21 papers evaluated, 14 included tools containing at least one propositional feature (66.67%). The following



		Objective 2: Type of Information from Discourse (Presence marked by x; number of features in parentheses)						Objective 5: Psychometric properties and implementation (present/absent)		
Paper	*Total features on tool*	*Linguistic*	*Propositional*	*Macrostructural*	*Pragmatic*	*Nonverbal/ Paralinguistic*	*Not categorizable*	*Rater reliability*	*Validity data*	*Implementation*
Ballard et al., 1999	5	x (2)	x (1)	x (1)			x (1)			
Behrns et al., 2009	6	x (1)		x (1)	x (4)			x	x	
Casilio et al., 2019	27	x (13)	x (4)	x (1)	x (1)		x (6)	x	x	
Copeland, 1989	26				x (21)	x (4)	x (1)	x		
Cupit et al., 2010	4	x (1)	x (2)				x (1)			
Dalton et al., 2020	2	x (1)	x (1)							
Doyle et al., 1996	1		x (1)					x	x	
Harmon et al., 2016	9				x (7)		x (2)			
Herrmann, 1989	2*		x (1*)		x (1*)	x (*)		x	x	
Hickey & Rondeau, 2005	7		x (1)	x (1)	x (7)					
Hula et al., 2003	1	x (1)						x		
Jacobs, 2001	4	x (1)	x (1)	x (1)	x (1)			x		
Kim & Wright, 2020	1	x (1)						x	x	
Ross, 1999	1						x (1)	x	x	
Terrell & Ripich, 1989	4*		x (1)	x (1*)	x (1*)	x (1)				
Ulatowska et al., 1981	2		x (1)	x (1)				x	x	
Ulatowska et al., 1983a	2		x (1)	x (1)				x	x	
Ulatowska et al., 1983b	2		x (1)	x (1)				x	x	
Ulatowska et al., 2003	3			x (3)				x	x	
Ulatowska et al., 2013	2		x (1)	x (1)				x		
Webster & Morris, 2019	1		x (1)						x	

** = see Supplementary Table of study for more detail.*

	Linguistic	Propositional	Macrostructural	Pragmatic	Nonverbal/ Paralinguistic	Not categorizable
features	21	18	13	43	5	12
tools	8	14	11	8	3	6
overall % features across tools	18.75	16.07	11.61	38.39	4.46	10.71
overall % of tools including this level	16	28	22	16	6	12

FIGURE 2 Summary of findings in objective two and objective five.

are several examples of propositional features from the included papers.

Across several studies, Ulatowska et al. asked listeners to rate discourse for clarity of content, which the authors defined as a proxy for cohesion (Ulatowska et al., 1981, 2013; Ulatowska, Doyel et al., 1983; Ulatowska, Freedman-Stern et al., 1983). In linguistics, cohesion has been traditionally defined as the links that hold a speech together and give it meaning. Cohesive devices include grammatical links such as pronoun referencing, substitutions and ellipsis; lexical links such as commonly co-occurring lexical items; and utterance links such as the use of conjunctions at the border between utterances (Halliday & Hasan, 2013). Main concept analysis from Dalton et al. (2020), evaluated gist production or informativeness, based on a normative list of main concepts, which were coded for accuracy and completeness based on rules from Nicholas and Brookshire (1995). Terrell and Ripich's (1989) DAP proposed an analysis tool where clinicians could note the presence or absence of essential steps in procedural narratives. Ballard and Thompson (1999) had judges rate 'content' on a numerical scale pre- and post-treatment. Finally, multiple studies had listeners rate overall informativeness of the discourse, or the amount of information produced (e.g., Cupit et al., 2010; Doyle et al., 1996; Jacobs, 2001; Webster & Morris, 2019). While this scoping review categorised 'informativeness' as propositional in most instances, the rating of 'overall informativeness' might be perceived by listeners to include propositional information with local and global coherence, including logical thematic and temporal organisation. That is, informativeness could be categorised into either propositional or macrostructural/planning levels. Given the broadness of rating 'informativeness' in this way, it is difficult to predict how raters conceptualised this metric.

Multiple groups conducted investigations on outcome measures that did not neatly fit into one of the Dipper and colleague levels, but which B.C.S. and S.G.D. felt were most closely related to cohesion and/or semantic appropriateness and completeness, thus falling into the propositional level. Jacobs (2001) had listeners rate the 'effectiveness' of discourse; Cupit et al. (2010) had listeners rate the speaker's 'ability to transmit the message';' and Herrmann (1989) had listeners rate a speaker's ability to convey a 'spontaneous and comprehensible message'. Note that these listener ratings of 'effectiveness', 'ability to transmit the message',' and 'spontaneous and comprehensive message' may also reflect coherence, described later, and therefore could be mapped onto the macrostructural/planning level.

Macrostructure/Planning

Of the 21 papers evaluated, 11 included tools containing at least one macrostructural/planning feature (52.34%).

The following are examples of macrostructural/planning features included in the papers.

Ulatowska et al. (2003) were the only group to include more than one macrostructure feature per tool. The macrostructure/planning discourse level examines the structure, gist, story content, framing and coherence of the discourse. Ulatowska and colleagues' body of work began by having listeners rate 'content', which they defined as a proxy of coherence (Ulatowska et al., 1981; Ulatowska, Doyel et al., 1983; Ulatowska, Freedman-Stern et al., 1983). Coherence was defined as measuring the overall 'meaningfulness' of the text. Cohesion (a propositional metric) differs from coherence in that discourse can be internally cohesive (using accurate grammatical and lexical linkages) but be incoherent. After examining the specific statements used by Ulatowska and colleagues, it is likely that content ratings indexed features across multiple levels including coherence (e.g., 'Is the story accurate in terms of the stimulus material?'), sequencing (e.g., 'Does the sequence of events make sense?'), and informativeness at the pragmatic level (e.g., 'Is it complete in the sense that it does not leave out any necessary information?' or 'Could you follow this procedure?'). In later work (e.g., Ulatowska et al., 2003), listeners were asked to explicitly rate 'coherence'.

Ulatowska et al. (2003) also developed a numerical scale for rating 'emplotment', which they defined as 'the ability to express information about an event in a narrative structural form, including all elements of the story or scenario'. This numerical scale resembles specific aspects of story grammar analyses, aiming to evaluate temporal ordering of story elements. Further, Ulatowska et al. (2003) had a numerical scale that evaluated 'referencing'. Authors defined referencing as how well elements such as characters, locations or time were unambiguously signalled in the story, which relate to aspects of macrostructural planning.

Copeland (1989) developed a rating scale of speech function, which broadly evaluated the 'completeness' of the transmitted message. Completeness suggests a meaningful transmission, and thus was categorised as closely related to coherence. Terrell and Ripich (1989)'s DAP was created for clinicians to examine the presence/absence of several planning components during narratives, including the production of setting and episodes during a narrative (e.g., initiating event, plan, consequence), and evaluating presence/absence of essential steps during procedural narratives. These steps can be interpreted as reflecting 'main concepts' as well as some aspect of sequencing. The DAP also evaluated presence/absence judgements for conversational skills related to macrostructural/planning, including topic initiation, maintenance and shifting, as well as turn-taking. Finally, the DAP provided an overall rating of coherence at the end of the tool. Hickey and Rondeau (2005) also evaluated turn-taking and continuation of topics or topic maintenance during conversation. Ballard and Thompson (1999) had judges perceptually rate coherence on a numerical scale for pre- and post-treatment narratives.

Pragmatic

Finally, several tools considered the pragmatic level of discourse, meaning the relationship of context, interpersonal factors, interactional factors and influences on discourse from situational and external sources. Of the 21 papers evaluated, eight included tools containing at least one pragmatic feature (38%). The following are examples of pragmatic features included in the papers.

Several included papers had tools that focused on pragmatics in particular (e.g., Copeland, 1989, with 21 pragmatic features; Harmon et al., 2016 and Hickey & Rondeau, 2005, with seven pragmatic features each). Jacobs (2001), Hickey and Rondeau (2005), and Harmon et al. (2016) evaluated an interpersonal factor, which had listeners rate their own comfort level when listening to discourse samples from individuals with aphasia. Harmon et al. (2016) probed listeners' personal feelings about the communication, for example, impatience and listeners' overall perception of competence. The DAP rated overall 'speech arts', which the tool defines as responding, requesting and asserting information, which are largely interactional behaviours. Herrmann (1989) had listeners rate the speaker's communication strategy, which had to do with the perception of the subject's intent to communicate (e.g., 'seemed motivated'). This appeared to reflect an interpersonal interpretation by the rater. Behrns et al (2009) had listeners rate speakers across several continuums that reflected interpersonal factors, such as difficult/easy to understand and interesting/not interesting. Cupit et al. (2010) had listeners rate 'degree of ease' in retelling narrative, which reflected an interpersonal factor, but which may also have fit in other categories. For example, 'degree of ease' may have reflected limited language access (linguistic component), limited cohesion (propositional component) and/or limited coherence (macrostructural component).

The APROCSA had one feature, 'overall communication impairment', which evaluated more holistic, message-level communication (Casilio et al., 2019). Similarly, Hickey and Rondeau (2005) and Ross and Wertz (1999) had listeners rate each speaker's 'overall communicative ability'. The DAP (Terrell & Ripich, 1989) similarly rated a speaker's overall communicative ability. 'Overall communicative ability' is quite difficult to categorise, as many pieces may contribute to it (e.g., linguistic, propositional, macrostructural, pragmatic), and it may be differently interpreted by each rater. Because of its overarching measurement, overall communicative ability' was categorised under pragmatics for the purposes of this scoping review. However, it is acknowledged that there are limitations on its ability to fit within a single category.

Monologic versus Dialogic

Fifteen studies with categorical scales and three studies using DME scales were appropriate for use with monologic discourse tasks (Tables 1 and 2). Of these, some were task-specific, in that they required a specific task in order to be used (e.g., core lexicon analysis, information unit and MCA). Other tools were appropriate for use across multiple genres. For example, the DAP (Terrell & Ripich, 1989) includes specific questions for different monologic discourse genres (e.g., procedural versus narrative). Eight studies with categorical scales and two studies using DME scales could be used to analyse dialogue.

Objective 3: Characteristics of speakers with aphasia across studies

The vast majority of studies that provided transcription-less tools included individuals with aphasia who were speakers of English, although one study included German speakers (Herrmann, 1989) and one included Swedish speakers (Behrns et al., 2009). While most studies included only individuals with chronic aphasia, several studies included participants across the recovery timeline, from acute (<6 months post-onset) to chronic (Doyle et al., 1996; Ross & Wertz, 1999; Ulatowska et al., 2003, 2013; Ulatowska, Doyel, et al., 1983; Ulatowska, Freedman-Stern, et al., 1983). The majority of studies used the WAB (Kertesz, 2007) for typing and severity determinations; however, three studies classified participants according to locus of damage (anterior, posterior, mixed; Ulatowska et al., 1981; Ulatowska, Doyel, et al., 1983; Ulatowska, Freedman-Stern, et al., 1983). One study used the BDAE (Goodglass & Kaplan, 1972) typing and severity system (Behrns et al., 2009), and Hula et al. (2003) used the Porch Index of Communicative Ability (Porch, 1967).

The extent to which each transcription-less scale evaluated discourse from participants with diverse aphasia types and severities was next evaluated. A total of 239 individuals with aphasia were included across all studies reviewed (note: Terrell & Ripich, 1989, did not report participant data). From the three studies that classified locus of damage (25 individuals total), 10 had anterior, nine posterior and six mixed damage (Ulatowska et al., 1981, 2013; Ulatowska, Doyel, et al., 1983; Ulatowska, Freedman-Stern, et al., 1983). For those studies which reported aphasia severity, 50 were described as mild, 29 as mild-moderate, 77 as moderate, one as moderate-severe, 24 as severe and three as very severe. It is important to note that most articles did not state how they defined severity categories, making inconsistencies within each group likely. For studies that used the WAB-R or BDAE to identify aphasia subtype, 51 individuals were classified as anomic, 20 as conduction, 11 as Wernicke's, five as transcortical sensory, 66 as Broca's, six as transcortical motor, 12 as global and one as unspecified/unclassifiable. In addition, 12 individuals with latent aphasia (defined as scoring above the WAB-R cutoff for aphasia but with continued complaints of language difficulties) were included. It is possible that some individuals classified as anomic in these studies may have more accurately been categorised as latent (for example, in Doyle et al., 1996). See Supplementary Materials for a more detailed description of participants with aphasia in each study, including any available demographic information.

Objective 4: Characteristics of raters/tool users

The included studies employed a diverse variety of raters/tool users. These ranged from naive listeners to students of speech pathology ('learners'), to the research team or clinical SLPs ('experts'). The definition of a naive listener varied across studies but was generally considered to be individuals with no training or coursework on aphasia or experience communicating with individuals with aphasia. Within this naive group, sometimes speech-language pathology students were explicitly excluded while other times SLP students were included if they had not taken coursework or worked with an individual with aphasia (Behrns et al., 2009; Cupit et al., 2010; Doyle et al., 1996; Hickey & Rondeau, 2005; Jacobs, 2001; Webster & Morris, 2019).

Five studies used learners as raters, ranging from undergraduate (Harmon et al., 2016; Kim & Wright, 2020) to master's level (Ballard & Thompson, 1999; Casilio et al., 2019; Harmon et al., 2016; Hickey & Rondeau, 2005; Ross & Wertz, 1999), and PhD students (Kim & Wright, 2020; Ulatowska et al., 1981). Seven studies used experts as raters, although raters were typically not familiar with the specific individuals with aphasia included in the study. In five studies, experts were certified SLPs (Copeland, 1989; Cupit et al., 2010; Hickey & Rondeau, 2005; Ulatowska, Doyel et al., 1983; Ulatowska, Freedman-Stern et al., 1983) and in two studies, experts were research team members (Casilio et al., 2019; Dalton et al., 2020). Other studies likely used expert research team members as raters, but this was not clearly reported (Herrmann, 1989; Ulatowska et al., 2003, 2013). One study used a mixed group of expert and learner raters, including two SLPs who were doctoral students, an experienced psychologist and one master's student (Hula et al., 2003). A final study did not include empirical data in their report but stated that the measure was designed for use by SLPs or anyone interested in the discourse of individuals with aphasia (Terrell & Ripich, 1989). See Supplementary Materials for a more detailed description of raters included in each study.

Objective 5: Psychometric data

See Figure 2 for a visualisation of whether each study evaluated reliability, validity and implementation. No formal evaluation of the quality of psychometric properties was conducted for the purposes of this scoping review. Instead, this scoping review documents how the original paper authors described the psychometric properties.

Rater reliability

More than half of the articles ($n = 14$) evaluated intra- and/or inter-rater reliability of the tool (see Supplementary materials of each study for detailed reporting). Included papers generally cited the finding as supportive of reliability.

Validity

Only two studies (Casilio et al., 2019; Kim & Wright, 2020) explicitly evaluated validity, and both studies demonstrated good concurrent validity. Casilio et al. (2019) demonstrated concurrent validity between the APROCSA and other perceptual connected speech tools (e.g., the spontaneous speech Fluency scale from the WAB-R). Kim and Wright (2020) demonstrated concurrent validity between core lexicon analysis and other micro- and macrolinguistic discourse measures extracted from transcriptions. After thoroughly reviewing each article, approximately half of the included studies ($n = 11$) were found to evaluate some aspect of validity but did not explicitly refer to these analyses as validity related. A large portion of these analyses correlated the transcription-less tool findings with transcription-derived metrics of discourse, with most included papers reporting that the transcription-less features were related to the transcription-derived features. Note that a variety of transcription-derived features were used, making it difficult to draw wider conclusions about the strength or types of relationships shown.

Implementation

None of the included studies piloted the tool in a clinical (non-research) setting, making the evaluation of implementability difficult. While some studies discussed using the scale in clinical settings (e.g., Terrell & Ripich, 1989), no empirical data of the tool's use in such a setting was provided. Further, none of the studies that included expert raters gathered feedback from raters about the potential of the tool to be implemented into a clinical setting. Few studies provided enough detail about the rating procedure to glean whether the tool could be used to score behaviours 'online', that is, during a session. Therefore, empirical work evaluating implementation was found to be lacking.

DISCUSSION

Several key findings and future directions were established through this scoping review. The following discussion of the findings is organised by objective.

Objectives 1 and 2: Rating tools across discourse levels and types

Some tools were specific to discourse type (e.g., dialogue), but many had potential to generalise across monologue and dialogue and across genres/tasks. It is therefore important for the tool user (e.g., SLP) to pick the most appropriate scale that relates to the discourse type they've chosen, the genre or task and the client's goals (e.g., ensure that the scale evaluates sequencing because this is the client's desired focus). Selection of the most appropriate discourse genre and task improves sensitivity to treatment changes and interpretation of results (Stark, 2019; Stark & Fukuyama, 2021). For example, syntax lacks test-retest reliability for the monologic procedural narrative of 'how to make a sandwich', making analysis of syntax from this task a poor choice to demonstrate syntactically-focused treatment change (Stark et al., 2023). However, other variables, like lexical-semantics and those related to motor output, appear to have more robust test-retest reliability across genres and tasks (Stark et al., 2023). Beyond test-retest reliability, one must also consider the appropriateness of the discourse metric as it relates to the task. If a clinician is interested in identifying an impairment specific to verbs, tasks that more heavily rely on spatial words and sequences (like moving from past to future) may be most appropriate.[1]

An evaluation of Figure 2 shows that the tools meeting the scoping review parameters tend to evaluate propositional and macrostructural levels most, with fewer evaluating the linguistic and pragmatic levels, and very few evaluating non-verbal or paralinguistic behaviours. Given that most transcription-based analyses evaluate linguistic metrics (as demonstrated by Bryant et al., 2016), this suggests that transcription-less tools are aiming to measure complementary though expanded aspects of discourse. Additionally, three tools included features across four levels (Casilio et al., 2019; Jacobs, 2001; Terrell & Ripich, 1989), three tools included features across three levels (Ballard & Thompson, 1999; Behrns et al., 2009; Herrmann, 1989), and most tools included features across only one or two levels (Copeland, 1989; Cupit et al., 2010; Dalton et al., 2020; Doyle et al., 1996; Harmon et al., 2016; Hickey & Rondeau, 2005; Kim & Wright, 2020; Ross & Wertz, 1999; Ulatowska et al., 1981, 1983a, 1983b, 2003, 2013; Webster & Morris, 2019). Even if the tool evaluated several levels, this did not mean that the evaluation within a level was exhaustive or comprehensive. For example, Jacobs (2001) provided perceptual rating opportunities across four levels, with only one feature per level: grammar (linguistic), informativeness (propositional), effectiveness (macrostructural) and listener's comfort level (pragmatic). While it is ideal for tools to evaluate across levels, limited features within a level provide limited opportunities for comprehensive evaluation and also may not be sensitive to certain aphasia profiles (e.g., grammar is often unaffected in those with milder aphasia).

There are numerous benefits to conducting analyses across levels. For example, Marini et al. (2011) conducted a multi-level analysis on narratives in Italian speakers with aphasia. Note that whilst Marini et al. (2011) relied upon transcriptions, this work provides a helpful schema for envisioning the benefits of multi-level analysis. Their analysis focused on three linguistic features (productivity, lexical processing, grammatical structuring), two proposition features (informativeness and cohesion), and two macrostructural/planning-level features (coherence and organisation). They reported that lexical and grammatical impairments led to decreased cohesion in their narrative samples, thus demonstrating how linguistic level impairments impact propositional level production. They also found that reduced levels of lexical informativeness were related to reduced coherence, thus relating a propositional level impairment to an impairment at the macrostructural/planning level. If they had focused on only a single level of discourse, these relationships between discourse levels may not have been realised.

The interconnectedness of discourse should be considered during treatment planning and progress monitoring. An individual presenting with deficits across multiple discourse levels may also demonstrate improvements across multiple levels, even if the treatment focus is predominantly on a single level. This also raises important concerns about selecting the most appropriate discourse level to target in therapy. Previous research has demonstrated that focussing therapy on more complex syntactic structures (Thompson et al., 2003), or training less typical members of a semantic category (Kiran & Thompson, 2003), lead to generalisation of simpler syntactic structures or more typical semantic category members, but the reverse is not true. While this complexity account has not been investigated with respect to discourse therapy, it would be prudent to do so. Current evidence in favour of the complexity account has examined complexity within a language domain (e.g., syntax or semantics), but Marini and colleagues' findings highlight the interrelatedness of discourse behaviours across language domains. Understanding whether the benefits of addressing more complex behaviours apply across language domains would aid SLPs in selecting the most appropriate treatment and treatment focus. Further, information from discourse when sampled across levels may better inform treatment focused on improving functional communication. For example, it may be the case that an individual wants to work on their sequencing and/or pragmatics, thus making it important to measure macrostructural/planning and pragmatic components of discourse. While the choices for evaluating specific discourse levels reflect the particular interests of the developers of each tool, given the benefits of multi-level analysis, expanding scales to have at least the option of assessing across discourse levels would be beneficial for a more comprehensive understanding of discourse in aphasia.

Marini et al. (2011) used transcriptions for their project, but there are several ways in which researchers and clinicians can conduct multi-level analysis using transcription-less tools. If evaluating a monologue sample, multiple transcription-less tools could be used to evaluate the same sample for different levels/features. For example, core lexicon analysis could be employed to analyse a linguistic feature; main concept analysis to analyse propositional features; story grammar analyses (e.g., Ulatowska and colleagues' scales on global structure, emplotment) to analyse macrostructural/planning features; and a rating of the speaker's 'overall communication' (e.g., Hickey & Rondeau, 2005) or a more specific probe of a listener's interpretation of speaker competency or clarity (e.g., Harmon et al., 2016) to analyse pragmatic features. However, there are drawbacks when combining many tools together to evaluate a single sample, among the greatest of which is the differing reliability and validity of those tools (see Discussion of Objective 5, psychometric properties).

Developing tools that evaluate a variety of levels with several features per level is an ideal starting point. From this psychometric perspective, the APROCSA (Casilio et al., 2019) is arguably the most well-validated and reliable perceptual tool available at present to evaluate multiple discourse levels. While it is specifically designed to evaluate linguistic features (~48% of tool features are linguistic), it enables the user to evaluate some propositional features (~15% of features), one global measure of macrostructure ('off topic', constituting ~4% of features), and includes at least one category ('overall communication impairment', ~4% of features) related to the pragmatic level. It also provides some information about speech (e.g., motor speech and fluency, ~22% of features), though there are no features that evaluate the non-verbal or paralinguistic level. Increasing the number of higher-level features (i.e., macrostructure, pragmatics), including non-verbal and/or paralinguistic features, and evaluating the tool's clinical implementability, would be promising next steps for encouraging multilevel discourse analysis in clinical practice. However, as discussed previously, it is incumbent upon SLPs to select the most appropriate analysis tool for their client and their client's outcomes rather than adhere to a single tool. Further development of transcription-less tools is therefore needed to ensure that SLPs are able to do so.

While not the main focus of this scoping review, most tools (Ballard & Thompson, 1999; Behrns et al., 2009; Casilio et al., 2019; Copeland, 1989; Cupit et al., 2010; Dalton et al., 2020; Doyle et al., 1996; Harmon et al., 2016; Hickey & Rondeau, 2005; Hula et al., 2003; Jacobs, 2001; Ross & Wertz, 1999; Ulatowska et al., 2003; Ulatowska et al., 2013; Ulatowska, Doyel, et al., 1983; Ulatowska, Freedman-Stern, et al., 1983; Webster & Morris, 2019) did not explicitly evaluate non-verbal and/or paralinguistic behaviours. Ulatowska et al. (2013) discussed collecting written reports of gestural communication accompanying speech, but these data were not evaluated by their rating scales. Only three studies explicitly evaluated non-verbal and/or paralinguistic behaviour. Herrmann et al.'s Communication Strategy Rating scale (1989) included three score categories that described using non-verbal or paralinguistic behaviour: Category 5 = 'Patient indicates by verbal or non-verbal means that he is unable to comprehend the questions'; Category 6 = 'Patient asks verbally or non-verbally for support when he fails to comprehend'; and Category 7 = 'Patient spontaneously employs compensatory (para- and non-verbal communicative) behaviour when communication by linguistic means cannot be established'. The DAP (Terrell & Ripich, 1989) includes a section which instructs the clinician to rate discourse abilities, including paralinguistic behaviour (e.g., stress, intonation, rate) and nonlinguistic behaviour (e.g., eye contact, gesture) on a Likert-style scale from poor to excellent based on performance in procedural, narrative and conversational tasks. The tool also has the clinician circle the paralinguistic (stress, intonation, rate) and non-linguistic behaviours (eye contact, gestures) which are demonstrated, thus providing qualitative data to complement the quantitative. Copeland (1989) had raters use two visual analogue scales to assess the amount and helpfulness of gesture, vocal inflection, facial expression and body movement. As noted in Tables 1 and 2, a majority of these transcription-less tools evaluated audio only (many of the DME scales were audio only in the research design), so it is not surprising that the tools did not include non-verbal features. Inclusion of multimodal data wherever possible is a key future direction, especially given the mounting evidence related to non-verbal behaviour being important for demonstrating communicative competence in aphasia (e.g., gesture: Akhavan et al., 2018; Cocks et al., 2011, 2018; Kong et al., 2015; Pritchard et al., 2013, 2015; van Nispen et al., 2017).

Objective 3: Inclusion of different aphasia types and severities

A strength of the studies reviewed here is that the majority included individuals across the range of aphasia severity and types. This provides assurance that transcription-less rating can be used to evaluate discourse from individuals across the severity spectrum. Further, there was an overarching focus of all tools to evaluate individuals with aphasia who were in the chronic stage, although some studies include individuals in the acute and sub-acute stage.

Two studies included individuals with latent aphasia (Dalton et al., 2020; Doyle et al., 1996), although Doyle and colleagues considered these individuals to have 'anomic' aphasia despite scoring above the standardised battery's cutoff for presence of aphasia. No study directly investigated the utility of transcription-less listener ratings in only individuals with latent aphasia. Individuals with latent aphasia are classically underserved because their language (and discourse) impairments are subtle—yet studies have also noted that they have residual impairments of language and that discourse is the best way to appreciate these (e.g., Fromm et al., 2017). Standardised batteries and isolated assessments of language are not sensitive enough to capture impairment and/or change post-therapy in this population. In the US system of rehabilitation, continued demonstration of need ensures that third-party payers (e.g., insurance) continue to support and reimburse for services. Therefore, sensitive assessments are extremely important. Discourse analysis may be one such assessment for those with latent aphasia. On the other side of the spectrum, the lack of tools evaluating non-verbal skills during discourse may adversely impact the tool's effectiveness for evaluating competence in those with severe aphasia. Listener ratings may be an excellent way to demonstrate both the need for therapy for individuals at the most extreme ends of the aphasia spectrum, as well as changes in communication as a result of therapy.

Objective 4: Inclusion of different rater types

While some scales have been specifically developed for naive users, others were developed for expert users. This is an important consideration when moving toward implementation. It is likely that scales developed for naive raters could be used by more experienced raters, but the reverse may not be true depending on the complexity of the concepts to be rated and the adequacy of definitions of these concepts. For example, in some scales, terms such as 'cohesion' and 'coherence' are not defined for raters. SLPs generally have some background in linguistics and/or language science and may be comfortable using these terms. However, this will not always be the case. Such a barrier could be easily alleviated by expanding instructions to include definitions for all terms. Definitions are important for terms that may be unfamiliar (like 'coherence' and 'cohesion') and for terms which have multiple operational definitions (such as 'discourse').

Furthermore, some consideration should be given towards whether the treating clinician is an acceptable rater of treatment outcome, since those studies which included SLPs ensured they were unfamiliar with the individuals with aphasia. If listener ratings are to be implemented in clinical practice, it is most likely that the treating clinician will complete the ratings. Thus, that clinician will have some familiarity with the speaking pattern of the individual with aphasia whose discourse is being evaluated. Alternatively, a rater might be a spouse or caregiver. These individuals are likely more familiar with the individual with aphasia, and likely aware of the focus and goal of therapy, which may impact reliability and validity. With the exception of trichotomous rating scales which may be more resistant to reliability issues (Metu et al., 2023), familiar listeners who are aware of therapy goals (e.g., clinicians, caregivers) may under- or over-estimate changes in response to therapy. This concern is supported by Hickey and Rondeau (2005) who reported different magnitudes of treatment change based on group (SLPs had the lowest magnitude change, naive listeners had the highest). Therefore, key future directions involve inclusion of familiar listeners to evaluate tool utility.

It is also important to consider the goal of therapy when identifying the most appropriate rater. In some instances (such as communication partner training) it may be most appropriate to have a familiar partner complete ratings (or even the individual with aphasia themselves, depending upon awareness of deficits). In other instances, such as if the goal of treatment is to improve a specific aspect of phonology, morphology, semantics, or syntax, the SLP may be the most qualified rater given their education and experience. It is also important to consider the extent to which intervention is resulting in a functional change outside the therapeutic context, and the extent to which that change is noticeable by individuals who are not as familiar with the client and who do not have knowledge of the client's goals. In this context, naive listeners who might encounter the client out in the community may be the best raters of treatment change. Unfamiliar listener communication situations are often the ones people with aphasia report as being most challenging, since unfamiliar listeners have less intrinsic motivation to communicate with a client than the treating SLP or familiar communication partners. If unfamiliar communication partners do not perceive a change following therapy, one may question whether one has met the overarching purpose of therapy.

The concept of most appropriate rater also relates back to reliability and validity: scale creators must demonstrate strong psychometric properties across different raters if indeed they believe the scale to be implementable by a variety of raters. This is elaborated on in the next section. Further, exploration of early implementation of these scales, such as by evaluating feasibility and utility in treatment settings with real clients, will lend valuable evidence for their use outside of a research-controlled environment.

Objective 5: Psychometric data

When scale developers are guided by Classical Test Theory, which is a theory of testing based around comparing an observed score on some assessment to a 'true' and 'error' score on the same assessment, the establishment of 'true' scores via reliability and validity investigations is best practice. A scale with strong psychometric properties also increases the confidence in using these scales to measure change post-therapy and/or change over time, such as evaluating differences from acute to chronic stages.

The majority of studies reported some form of rater reliability, with the majority evaluating inter-rater, rather than intra-rater, reliability. This is interesting, given intra-rater reliability is likely going to be the most important psychometric property of rater reliability to establish for tools in clinical settings. This is because the tool is likely to be used repeatedly by the same, single provider (e.g., SLP), for example to show therapy-induced change. Studies used a variety of statistics including correlation, intraclass correlation coefficient and percentage agreement to quantify agreement magnitude or absolute agreement. While no formal assessment of the quality of these analyses was conducted for the purposes of this scoping review, original paper authors cited rater reliability of their tools as 'acceptable' or similar.

A clear next step is the evaluation of each tool's test-retest reliability, to ensure that the tool has utility in longitudinal examinations, that is, that it can reliably measure change over time or change due to therapy. Test-retest studies on language produced during discourse have highlighted the complex relationship between reliability and aphasia severity, task, and words produced (Stark et al., 2023). Studies suggest that, minimally, tools must be evaluated for their test-retest reliability across different discourse genres and take into account the known variability related to aphasia severity and type, vocabulary access and length of sample.

Finally, of particular interest to this scoping review, which has focused on the distinctions between different language (and non-verbal/paralinguistic) components to discourse, is the extent to which reliability of a tool extends to each feature of the tool. For example, are linguistic features more reliably rated than pragmatic features? Is one linguistic feature more reliably rated than another linguistic feature? An argument might be that pragmatic features involve more rater 'interpretation' than linguistic behaviours, leading to less reliability. Indeed, the APROCSA (Casilio et al., 2019) highlights the need to conduct feature-level reliability analyses and was one of the few studies to acknowledge the potential of rater agreement to vary by feature. In Casilio et al. (2019) Figure 1, intraclass correlation coefficients (a metric of rater reliability) varied by feature for expert and learner raters (graduate clinicians in speech-language pathology). For example, the intraclass correlation coefficient for phonemic paraphasias (a linguistic feature) was relatively poor for both expert and student raters, suggesting that this feature, in particular, was difficult to reliably rate perceptually. Yet, another linguistic and related measure, neologisms, were rated reliably by both the expert and learner raters. A propositional feature, 'meaning unclear', had moderate reliability for both experts and learners, while a macrostructural feature, 'off topic', had poor rater reliability for both groups. 'Overall communication impairment', categorised as a pragmatic measure for the purposes of this scoping review, had very strong rater reliability in both groups. Thanks to the robust psychometric evaluation of rater reliability at the feature level, Casilio et al. (2019) illustrate the need for other studies to replicate this type of analysis. Ultimately, tools should be refined to include features that are reliably rated by the raters of interest (e.g., experts versus naive), and to exclude (or make 'optional') features that are less reliably rated.

Half of the transcription-less rating scales reported on validity. Concurrent validity was most commonly evaluated, typically by correlating the scores from the transcription-less tools with logically-related scores derived from transcripts (e.g., 'correct information units', Nicholas & Brookshire, 1993, should be related to perceptual ratings of informativeness). Minimally, psychometric data including intra- and inter-rater reliability, test-retest reliability and construct validity should be available prior to widespread use of a rating scale (Pritchard et al., 2018). Ideally, face and ecological validity should also be established to confirm the utility of any rating scale. Because of the potential to use these tools to sensitively demonstrate language impairment in very subtle or very severe aphasia, 'known groups' validity is also ideal to evaluate. For example, a tool with strong known groups validity would be able to sensitively differentiate individuals with severe aphasia from those with moderate aphasia. Identifying known groups validity strengthens the evidence for the tool's usefulness in sensitively and specifically identifying impairments (and strengths) across aphasia presentations.

It should be noted that even if tools have demonstrated validity between transcription-less and transcription-derived metrics, this does not mean that transcription-less methods are completely comparable to a full, transcription-derived discourse analysis. As discussed in the Introduction, transcription enables detailed analysis that transcription-less can likely not replicate. However, the draw to transcription-less analysis is its potential for implementation and uptake in the clinical profession, enabling an overview of discourse strengths and weaknesses that can further guide treatment or more detailed assessment.

Finally, as laid out in the Introduction, a prime goal of transcription-less analyses is to improve the implementation of discourse analysis into clinical practice. Cruice et al. (2020) found that a large number of their clinical SLP respondents would implement discourse analysis if they could do so within 60 min. Given that most SLP sessions last between 45–60 min in the United States (Cavanaugh et al., 2021), a tool that enables online, reliable scoring of discourse behaviours would fulfil this need. Unfortunately, no study in this scoping review garnered feedback from expert raters about implementation potential, and no study implemented the tool in a clinical setting. Therefore, even whilst most tools in this scoping review provide good evidence for rater reliability and burgeoning evidence for tool validity (especially concurrent with some transcription-derived measures), the lack of empirical investigation of implementation continues to limit the use of these tools in clinical practice. This is coupled with the statistics discussed in the Introduction, which also suggest that clinician training about discourse (and the complexity of evaluating discourse) are important factors influencing the lack of discourse analysis in clinical practice. As such, there remains a wide research-to-practice gap as it relates to using transcription-less tools, despite transcription-less tools' potential to alleviate many barriers to discourse analysis.

Limitations

There are several limitations to acknowledge.

The scoping review may have been limited in its search terms, obtaining more monologue than dialogue examples of transcription-less tools. The rationale for search term selection was that 'connected speech' and 'spontaneous speech' would capture monologue and dialogue, but in hindsight, additional specific search terms such as 'interview' or 'conversation' may have increased the number of papers explicitly evaluating dialogue. The authors also acknowledge a risk of bias in that B.C.S. reviewed all abstracts, and then both S.G.D. and B.C.S. reviewed all papers.

A particular limitation for objective two (levels of discourse being evaluated) was the difficulty differentiating into which level some of the discourse behaviours fell. As discussed in the results, it is difficult to predict how raters conceptualise metrics that aim to measure some comprehensive aspect of the speech or communication ability, that is, 'overall informativeness', and 'overall communication impairment'. BCS and SGD categorised these to the best of their abilities but recognise the difficulty in assigning some of these behaviour classifications into a single category.

Future directions for transcription-less tool creation and validation

This scoping review suggests some similarities between transcription-less and transcription approaches, at least for linguistic and propositional information (e.g., Armstrong et al., 2007; Bryant et al., 2019; Kong & Wong, 2018; Ruiter et al., 2022). However, not all studies included in this scoping review directly investigated the relationship between transcription-less and transcription analyses, and there are unique reasons for choosing to do transcription or transcription-less analyses. A logical next step is the direct comparison of transcription and transcription-less analyses rated by SLPs, ideally taking into account feasibility and utility to demonstrate implementation potential and construct and face validity.

An important consideration for any assessment is the risk of bias inherent in the method. There is a body of research demonstrating that standardised assessments are often biased or yield less valid results for minoritised populations (e.g., Lynch & Davison, 2022; Molrine & Pierce, 2002; Olbert et al., 2018) The reasons for this are numerous but include lack of diverse representatives during test development and in normative samples. This is a particular concern for aphasiologists since stroke demographics demonstrate that African American and Hispanic individuals experience higher rates of stroke and more negative outcomes (Acton et al., 2022; Centers for Disease Control and Prevention (CDC), 2005; M. Jacobs & Ellis, 2022) than their white counterparts. Biased assessment measures can lead to the under- or over-diagnosis of disorders in individuals (e.g., over-diagnosis of schizophrenia in African American males; Olbert et al., 2018), and likely contribute to the disparities in outcomes experienced by individuals of colour following a stroke. An issue that arose whilst conducting this review was the limited information provided about the raters (i.e., listeners). For example, no study provided information on the raters' race/ethnicity or additional languages spoken (beyond English) or language status (i.e., bilingualism). This omission of information does not enable readers to ensure that people from minoritised populations are factored into tool development, nor does it ensure that validation of the tool is actually inclusive of all tool users. It was also noted that very little information was provided regarding race/ethnicity, languages spoken and language status for the individuals with aphasia within each study. This finding has been reported elsewhere in relation to an omission of this critical information from clinical trials in aphasia (Nguy et al., 2022), and future research must improve inclusivity of both raters and participants to improve validity of findings. While listener perceptions of communication are not free from risk of bias, it may be possible to mediate these risks through more diverse involvement of raters and participants, specific training in the use of the tools, and by improving graduate education programs in speech pathology to ensure the development of culturally competent clinicians.

Despite a recent push to integrate lived experiences into aphasia assessments and treatments (Hinckley, 2023), this scoping review did not identify that any transcription-less scales were developed in concert with partners (i.e., clients with aphasia, practising SLPs). It is well known that co-design elicits ideas and fosters an environment in which concepts, knowledge and lived experiences are applied to develop tools that meet individuals' needs, which are then more valid, have higher utility, take into account sociocultural considerations and are more likely to be adopted for use (Page et al., 2016; Sanders & Stappers, 2008; Wilson et al., 2015). Additionally, co-design procedures that specifically include minoritised individuals demonstrate less bias than traditionally designed measures (Olbert et al., 2018), suggesting that co-design of transcription-less discourse measures may reduce concerns of listener bias in ratings. The alternative to co-design is to risk research waste, where time, money and effort are invested in materials or tools that will not be implemented (Page et al., 2016; Sanders & Stappers, 2008). Further, because of the lack of partner involvement, it is unclear if transcription-less tools are measuring exactly what SLPs need and want to measure, such as features that directly relate to functional and person-centred outcomes. The development (or refinement) of transcription-less tools as being partner-driven is a tangible and necessary future direction.

Conclusions

This scoping review highlights numerous tools for perceptually analysing discourse without transcription, specific to individuals with aphasia. These tools enable analysis across several discourse levels, with few tools having multiple features per level, and there are more tools developed for analysing monologue than dialogue. A variety of raters have been used to 'test' these tools, and consideration should be given to the implications of employing 'naive', 'learner', and 'expert' raters, depending on the desired outcome of the tool. The tools tended to focus on discourse in chronic aphasia, with clear future directions for assessing whether the tools might be used to evaluate discourse from acute to chronic stages, or indeed, for other longitudinal means (e.g., pre/post treatment, over time). The review suggests that these tools have burgeoning psychometric properties, especially around rater reliability and validity, but considerable ground to gain related to enhanced testing of reliability (e.g., evaluating test-retest reliability; evaluating reliability for each specific function evaluated by the tool), validity (e.g., more discriminant and known groups validity studies), and importantly, implementation in clinical settings.

ACKNOWLEDGEMENT

No funding was received in support of this manuscript.

CONFLICT OF INTEREST STATEMENT

The authors have no financial or non-financial conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study. Data extracted are available in supplementary materials.

ORCID

Sarah Grace Dalton https://orcid.org/0000-0002-1504-8002

ENDNOTE

[1] For excellent guidance on selecting the appropriate discourse types or tasks for a client, Leaman and Archer's (2023) recent tutorial may be helpful when paired with this review to select an appropriate transcription-less tool.

REFERENCES

Acton, E.K., Abbasi, M.H. & Kasner, S.E. (2022) Evaluating age, sex, racial, and ethnic representation in acute ischemic stroke trials, 2010 to 2020: a systematic review and meta-analysis. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 11(8), e024651. https://doi.org/10.1161/JAHA.121.024651

Akhavan, N., Göksun, T. & Nozari, N. (2018) Integrity and function of gestures in aphasia. *Aphasiology*, 32(11), 1310–1335. https://doi.org/10.1080/02687038.2017.1396573

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

Armstrong, L., Brady, M., Mackenzie, C. & Norrie, J. (2007) Transcription-less analysis of aphasic discourse: a clinician's dream or a possibility? *Aphasiology*, 21(3–4), 355–374. https://doi.org/10.1080/02687030600911310

Ballard, K.J. & Thompson, C.K. (1999) Treatment and generalization of complex sentence production in agrammatism. *Journal of Speech, Language, and Hearing Research*, 42(3), 690–707. https://doi.org/10.1044/jslhr.4203.690

Behrns, I., Wengelin, A., Broberg, M. & Hartelius, L. (2009) A comparison between written and spoken narratives in aphasia. *Clinical Linguistics & Phonetics*, 23(7), 507–528. https://doi.org/10.1080/02699200902916129

Berube, S., Nonnemacher, J., Demsky, C., Glenn, S., Saxena, S., Wright, A., Tippett, D.C. & Hillis, A.E. (2019) Stealing cookies in the twenty-first century: measures of spoken narrative in healthy versus speakers with aphasia. *American Journal of Speech-Language Pathology*, 28(1S), 321–329. https://doi.org/10.1044/2018_AJSLP-17-0131

Boles, L. (1998) Conversational discourse analysis as a method for evaluating progress in aphasia: a case report. *Journal of Communication Disorders*, 31, 261–274.

Boyle, M. (2011) Discourse treatment for word retrieval impairment in aphasia: the story so far. *Aphasiology*, 25(11), 1308–1326. https://doi.org/10.1080/02687038.2011.596185

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

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

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

Bryant, L., Ferguson, A., Valentine, M. & Spencer, E. (2019) Implementation of discourse analysis in aphasia: investigating the feasibility of a Knowledge-to-Action intervention. *Aphasiology*, 33(1), 31–57. https://doi.org/10.1080/02687038.2018.1454886

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

Casilio, M., Rising, K., Beeson, P.M., Bunton, K. & Wilson, S.M. (2019) Auditory-perceptual rating of connected speech in aphasia. *American Journal of Speech-Language Pathology*, 28(2), 550–568. https://doi.org/10.1044/2018_AJSLP-18-0192

Cavanaugh, R., Kravetz, C., Jarold, L., Quique, Y., Turner, R. & Evans, W.S. (2021) Is there a research–practice dosage gap in aphasia rehabilitation? *American Journal of Speech-Language Pathology*, 30(5), 2115–2129. https://doi.org/10.1044/2021_AJSLP-20-00257

Centers for Disease Control and Prevention (CDC). (2005) Differences in disability among black and white stroke survivors—United States, 2000–2001. *MMWR. Morbidity and Mortality Weekly Report*, 54(1), 3–6.

Clough, S. & Gordon, J.K. (2020) Fluent or nonfluent? Part A. Underlying contributors to categorical classifications of fluency in aphasia. *Aphasiology*, 34(5), 515–539. https://doi.org/10.1080/02687038.2020.1727709

Cocks, N., Byrne, S., Pritchard, M., Morgan, G. & Dipper, L. (2018) Integration of speech and gesture in aphasia. *International Journal of Language and Communication Disorders*, 53(3), 584–591. https://doi.org/10.1111/1460-6984.12372

Cocks, N., Dipper, L., Middleton, R. & Morgan, G. (2011) What can iconic gestures tell us about the language system? A case of conduction aphasia. *International Journal of Language and Communication Disorders*, 46(4), 423–436. https://doi.org/10.3109/13682822.2010.520813

Copeland, M. (1989) An assessment of natural conversation with broca's aphasics. *Aphasiology*, 3(4), 301–306. https://doi.org/10.1080/02687038908249001

Croteau, C., McMahon-Morin, P., Le Dorze, G., Power, E., Fortier-Blanc, J. & Davis, G.A. (2018) Exploration of a quantitative method for measuring behaviors in conversation. *Aphasiology*, 32(3), 247–263. https://doi.org/10.1080/02687038.2017.1350629

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), 417–442. https://doi.org/10.1111/1460-6984.12528

Cupit, J., Rochon, E., Leonard, C. & Laird, L. (2010) Social validation as a measure of improvement after aphasia treatment: its usefulness and influencing factors. *Aphasiology*, 24(11), 1486–1500. https://doi.org/10.1080/02687031003615235

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.H. & Richardson, J.D. (2019) A large-scale comparison of main concept production between persons with aphasia and persons without brain injury. *American Journal of Speech-Language Pathology*, 28(1S), 293–320. https://doi.org/10.1044/2018_AJSLP-17-0166

Dalton, S.G., Stark, B.C., Fromm, D., Apple, K., MacWhinney, B., Rensch, A. & Rowedder, M. (2022) Validation of an automated procedure for calculating core lexicon from transcripts. *Journal of Speech, Language, and Hearing Research*, 65(8), 2996–3003. https://doi.org/10.1044/2022_JSLHR-21-00473

Damico, J.S., Simmons-Mackie, N., Oelschlaeger, M., Elman, R. & Armstrong, E. (1999) Qualitative methods in aphasia research: basic issues. *Aphasiology*, 13(9–11), 651–665. https://doi.org/10.1080/026870399401768

de Beer, C., de Ruiter, J.P., Hielscher-Fastabend, M. & Hogrefe, K. (2019) The production of gesture and speech by people with aphasia: influence of communicative constraints. *Journal of Speech, Language, and Hearing Research*, 62(12), 4417–4432. https://doi.org/10.1044/2019_JSLHR-L-19-0020

de Beer, C., Hogrefe, K., Hielscher-Fastabend, M. & de Ruiter, J.P. (2020) Evaluating models of gesture and speech production for people with aphasia. *Cognitive Science*, 44(9), e12890. https://doi.org/10.1111/cogs.12890

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

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

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

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

Doyle, P.J., Tsironas, D., Goda, A.J. & Kalinyak, M. (1996) The relationship between objective measures and listeners' judgments of the communicative informativeness of the connected discourse of adults with aphasia. *American Journal of Speech-Language Pathology*, 5(3), 53–60. https://doi.org/10.1044/1058-0360.0503.53

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

Frederiksen, C.H., Bracewell, R.J., Breuleux, A. & Renaud, A. (1990) The cognitive representation and processing of discourse: function and dysfunction. In: Joanette, Y. & Brownell, H.H. (Eds.) *Discourse ability and brain damage: theoretical and empirical perspectives*. Springer, pp. 69–110. https://doi.org/10.1007/978-1-4612-3262-9_4

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

Goodglass, H. & Kaplan, E. (1972) *Boston diagnostic aphasia examination*. Lea & Febiger.

Goodman, S.N., Fanelli, D. & Ioannidis, J.P.A. (2016) What does research reproducibility mean? *Science Translational Medicine*, 8(341), 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Gordon, J.K. & Clough, S. (2020) How fluent? Part B. Underlying contributors to continuous measures of fluency in aphasia. *Aphasiology*, 34(5), 643–663. https://doi.org/10.1080/02687038.2020.1712586

Grimshaw, J. (2010) *A knowledge synthesis chapter*. Ottawa, Canada: Canadian Institutes of Health Research.

Halliday, M.A.K. (1985) *An introduction to functional grammar*, 1st edition, Edward Arnold.

Halliday, M.A.K. & Hasan, R. (2013) *Cohesion in english*. Routledge. https://doi.org/10.4324/9781315836010

Harmon, T.G., Jacks, A., Haley, K.L. & Faldowski, R.A. (2016) Listener perceptions of simulated fluent speech in nonfluent aphasia. *Aphasiology*, 30(8), 922–942. https://doi.org/10.1080/02687038.2015.1077925

Herrmann, M. (1989) Communicative skills in chronic and severe nonfluent aphasia*1. *Brain and Language*, 37(2), 339–352. https://doi.org/10.1016/0093-934X(89)90022-9

Hickey, E. & Rondeau, G. (2005) Social validation in aphasiology: does judges' knowledge of aphasiology matter? *Aphasiology*, 19(3–5), 389–398. https://doi.org/10.1080/02687030444000831

Hinckley, J. (2023) Stakeholder-Engaged Research: examples from Aphasia. *Topics in Language Disorders*, 43(1), 1. https://doi.org/10.1097/TLD.0000000000000307

Hula, W., McNeil, M., Doyle, P., Rubinsky, H. & Fossett, T. (2003) The inter-rater reliability of the story retell procedure. *Aphasiology*, 17(5), 523–528. https://doi.org/10.1080/02687030344000139

Jacobs, B.J. (2001) Social validity of changes in informativeness and efficiency of aphasic discourse following linguistic specific treatment (LST). *Brain and Language*, 78(1), 115–127. https://doi.org/10.1006/brln.2001.2452

Jacobs, M. & Ellis, C. (2022) Racial disparities in post-stroke aphasia: a need to look beyond the base analysis. *Journal of the National Medical Association*, https://doi.org/10.1016/j.jnma.2022.01.009

Kertesz, A. (2007) *Western aphasia battery—revised*. Pearson. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Speech-%26-Language/Western-Aphasia-Battery-Revised/p/100000194.html?tab=product-details

Kim, H. & Wright, H.H. (2020) Concurrent validity and reliability of the core lexicon measure as a measure of word retrieval ability in aphasia narratives. *American Journal of Speech-Language Pathology*, 29(1), 101–110. https://doi.org/10.1044/2019_AJSLP-19-0063

Kintsch, W. & van Dijk, T.A. (1978) Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kintz, S. & Wright, H.H. (2017) Discourse measurement in aphasia research. *Aphasiology*, 00(00), 1–3. https://doi.org/10.1080/02687038.2017.1398807

Kiran, S. & Thompson, C.K. (2003) The role of semantic complexity in treatment of naming deficits. *Journal of Speech, Language, and Hearing Research*, 46(4), 773–787. https://doi.org/10.1044/1092-4388(2003/061)

Kong, A.P.-H., Law, S.-P., Kwan, C.C.-Y., Lai, C. & Lam, V. (2015) A coding system with independent annotations of gesture forms and functions during verbal communication: development of a Database of Speech and GEsture (DoSaGE). *Journal of Nonverbal Behavior*, 39(1), 93–111. https://doi.org/10.1007/s10919-014-0200-6

Kong, A.P.-H. & Wong, C.W.-Y. (2018) An integrative analysis of spontaneous storytelling discourse in aphasia: relationship with listeners' rating and prediction of severity and fluency status of aphasia. *American Journal of Speech-Language Pathology*, 27(4), 1491–1505. https://doi.org/10.1044/2018_AJSLP-18-0015

Labov, W. (1972) *Language in the inner city: studies in the black english vernacular*. University of Pennsylvania Press.

Lanyon, L. & Rose, M.L. (2009) Do the hands have it? The facilitation effects of arm and hand gesture on word retrieval in aphasia. *Aphasiology*, 23(7–8), 809–822. https://doi.org/10.1080/02687030802642044

Leaman, M.C. & Archer, B. (2023) Choosing discourse types that align with person-centered goals in aphasia rehabilitation: a clinical tutorial. *Perspectives of the ASHA Special Interest Groups*, 8(2), 254–273. https://doi.org/10.1044/2023_PERSP-22-00160

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

Levelt, W. (1989) *Speaking: from intention to articulation*. MIT Press.

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

Lomas, J., Pickard, L., Bester, S., Elbard, H., Finlayson, A. & Zoghaib, C. (1989) The communicative effectiveness index: developmental and psychometric evaluation of a functional communication measure for adult aphasia. *Journal of Speech and Hearing Disorders*, 54(1), 113–124. https://doi.org/10.1044/jshd.5401.113

Lynch, A. & Davison, K. (2022) Gendered expectations on the recognition of ADHD in young women and educational implications. *Irish Educational Studies*, 0(0), 1–21. https://doi.org/10.1080/03323315.2022.2032264

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

Metu, J., Kotha, V. & Hillis, A.E. (2023) Evaluating fluency in aphasia: fluency scales, trichotomous judgements, or machine learning. *Aphasiology*, 1–13. https://doi.org/10.1080/02687038.2023.2171261

Molrine, C.J. & Pierce, R.S. (2002) Black and white adults' expressive language performance on three tests of aphasia. *American Journal of Speech-Language Pathology*, 11(2), 139–150. https://doi.org/10.1044/1058-0360(2002/014)

Nguy, B., Quique, Y.M., Cavanaugh, R. & Evans, W.S. (2022) Representation in aphasia research: an examination of u.s. treatment studies published between 2009 and 2019. *American Journal of Speech-Language Pathology*, 31(3), 1424–1430. https://doi.org/10.1044/2022_AJSLP-21-00269

Nicholas, L.E. & Brookshire, R.H. (1993) A System for Quantifying the Informativeness and Efficiency of the Connected Speech of Adults with Aphasia. *Journal of Speech, Language, and Hearing Research*, 36(2), 338–350. https://doi.org/10.1044/jshr.3602.338

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

Olbert, C.M., Nagendra, A. & Buck, B. (2018) Meta-analysis of Black vs. White racial disparity in schizophrenia diagnosis in the United States: do structured assessments attenuate racial disparities? *Journal of Abnormal Psychology*, 127(1), 104–115. https://doi.org/10.1037/abn0000309

Olness, G.S. & Ulatowska, H.K. (2011) Personal narratives in aphasia: coherence in the context of use. *Aphasiology*, 25(11), 1393–1413. https://doi.org/10.1080/02687038.2011.599365

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

Page, G.G., Wise, R.M., Lindenfeld, L., Moug, P., Hodgson, A., Wyborn, C. & Fazey, I. (2016) Co-designing transformation research: lessons learned from research on deliberate practices for transformation. *Current Opinion in Environmental Sustainability*, 20, 86–92. https://doi.org/10.1016/j.cosust.2016.09.001

Peters, M.D.J., Marnie, C., Tricco, A.C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C.M. & Khalil, H. (2020) Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis*, 18(10), 2119–2126. https://doi.org/10.11124/JBIES-20-00167

Porch, B. (1967) *Porch index of communicative ability: theory and development*. Consulting Psychologists Press.

Pritchard, M., Cocks, N. & Dipper, L. (2013) Iconic gesture in normal language and word searching conditions: a case of conduction aphasia. *International Journal of Speech-Language Pathology*, 15(5), 524–534. https://doi.org/10.3109/17549507.2012.712157

Pritchard, M., Dipper, L., Morgan, G. & Cocks, N. (2015) Language and iconic gesture use in procedural discourse by speakers with aphasia. *Aphasiology*, 29(7), 826–844. https://doi.org/10.1080/02687038.2014.993912

Pritchard, M., Hilari, K., Cocks, N. & Dipper, L. (2017) Reviewing the quality of discourse information measures in aphasia. *International Journal of Language & Communication Disorders*, 1–44. 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

Qin, Y., Lee, T. & Kong, A.P.H. (2020) Automatic assessment of speech impairment in cantonese-speaking people with aphasia. *IEEE Journal of Selected Topics in Signal Processing*, 14(2), 331–345. https://doi.org/10.1109/JSTSP.2019.2956371

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), Article 1. https://doi.org/10.3390/brainsci11010110

Ross, K.B. & Wertz, R. (1999) Comparison of impairment and disability measures for assessing severity of, and improvement in, aphasia. *Aphasiology*, 13(2), 113–124. https://doi.org/10.1080/026870399402235

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*, 1–20. https://doi.org/10.1080/02687038.2022.2109124

Rumelhart, D.E. (1975) Notes on a schema for stories. In: Bobrow, D.G. & Collins, A. (Eds.), *Representation and understanding*. Morgan Kaufmann, pp. 211–236. https://doi.org/10.1016/B978-0-12-108550-6.50013-6

Sanders, E.B.-N. & Stappers, P.J. (2008) Co-creation and the new landscapes of design. *CoDesign*, 4(1), 5–18. https://doi.org/10.1080/15710880701875068

Sekine, K. & Rose, M.L. (2013) The relationship of aphasia type and gesture production in people with aphasia. *American Journal of Speech-Language Pathology*, 22(4), 662–672. https://doi.org/10.1044/1058-0360(2013/12-0030)

Slobin, D.I. (1996) From “thought and language” to ’thinking for speaking. In: Gumperz, J.J. & Levinson, S.C. (Eds.), *Rethinking linguistic relativity*. Cambridge University Press, pp. 70–96.

Sperber, D. & Wilson, D. (1986) *Relevance: communication and cognition*, 1st edition. Blackwell Publishers.

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

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

Stark, B.C., Basilakos, A., Hickok, G., Rorden, C., Bonilha, L. & Fridriksson, J. (2019) Neural organization of speech production: a lesion-based study of error patterns in connected speech. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, 117, 228–246. https://doi.org/10.1016/j.cortex.2019.02.029

Stark, B.C., Dutta, M., Murray, L.L., Bryant, L., Fromm, D., MacWhinney, B., … & Sharma, S. (2021) Standardizing assessment of spoken discourse in aphasia: a working group with deliverables. *American Journal of Speech-Language Pathology*, 30(1S), 491–502. https://doi.org/10.1044/2020_AJSLP-19-00093

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

Stark, B.C. & Fukuyama, J. (2021) Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. *Language, Cognition and Neuroscience*, 36(5), 562–585. https://doi.org/10.1080/23273798.2020.1862258

Stein, N.L. & Glenn, C. (1979) An analysis of story comprehension in elementary school children: a test of a schema. In: Freedle, R.O. (Ed.), *New directions in discourse processing*. Ablex Publishing.

Stein, N.L. & Glenn, C.G. (1975) *An Analysis of Story Comprehension in Elementary School Children: A Test of a Schema*. https://eric.ed.gov/?id=ED121474

Swinburn, K., Porter, G. & Howard, D. (2004) *Comprehensive aphasia test*. Psychology Press.

Terrell, B. & Ripich, D. (1989) Discourse competence as a variable in intervention. *Seminars in Speech and Language*, 10(04), 282–297. https://doi.org/10.1055/s-2008-1064269

Thompson, C.K., Shapiro, L.P., Kiran, S. & Sobecks, J. (2003) The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia. *Journal of Speech, Language, and Hearing Research*, 46(3), 591–607. https://doi.org/10.1044/1092-4388(2003/047)

Trupe, E.H. (1984) Reliability of rating spontaneous speech in the western aphasia battery: implications for classification. *Clinical Aphasiology: Proceedings of the Conference 1984*. Clinical Aphasiology Conference.

Ulatowska, H. & Bond Chapman, S. (1989) Discourse considerations for aphasia management. *Seminars in Speech and Language*, 10(04), 298–314. https://doi.org/10.1055/s-2008-1064270

Ulatowska, H.K., Allard, L. & Chapman, S.B. (1990) Narrative and procedural discourse in aphasia. In: Joanette, Y. & Brownell, H.H. (Eds.), *Discourse ability and brain damage: theoretical and empirical perspectives*. Springer, pp. 180–198. https://doi.org/10.1007/978-1-4612-3262-9_8

Ulatowska, H.K., Doyel, A.W., Stern, R.F., Haynes, S.M. & North, A.J. (1983a) Production of procedural discourse in aphasia. *Brain and Language*, 18(2), 315–341. https://doi.org/10.1016/0093-934X(83)90023-8

Ulatowska, H.K., Freedman-Stern, R., Doyel, A.W., Macaluso-Haynes, S. & North, A.J. (1983b) Production of narrative discourse in aphasia. *Brain and Language*, 19(2), 317–334. https://doi.org/10.1016/0093-934X(83)90074-3

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

Ulatowska, H.K. & Olness, G.S. (2004) Discourse. In *The MIT encyclopedia of communication disorders* (pp. 300–302). MIT Press.

Ulatowska, H.K., Olness, G.S., Wertz, R.T., Thompson, J.L., Keebler, M.W., Hill, C.L. & Auther, L.L. (2001) Comparison of language impairment, functional communication, and discourse measures in African-American aphasic and normal adults. *Aphasiology*, 15(10–11), 1007–1016. https://doi.org/10.1080/02687040143000357

Ulatowska, H.K., Reyes, B., Santos, T.O., Garst, D., Vernon, J. & McArthur, J. (2013) Personal Narratives in Aphasia: understanding Narrative Competence. *Topics in Stroke Rehabilitation*, 20(1), 36–43. https://doi.org/10.1310/tsr2001-36

Ulatowska, H., Streit Olness, G., Wertz, R., Samson, A., Keebler, M. & Goins, K. (2003) Relationship between discourse and Western Aphasia Battery performance in African Americans with aphasia. *Aphasiology*, 17(5), 511–521. https://doi.org/10.1080/0268703034400102

van Dijk, T.A. & Kintsch, W. (1983) *Strategies of discourse comprehension*. Academic Press.

van Nispen, K., van de Sandt-Koenderman, M., Sekine, K., Krahmer, E. & Rose, M.L. (2017) Part of the message comes in gesture: how people with aphasia convey information in different gesture types as compared with information in their speech. *Aphasiology*, 31(9), 1078–1103. https://doi.org/10.1080/02687038.2017.1301368

Webster, J. & Morris, J. (2019) Communicative informativeness in aphasia: investigating the relationship between linguistic and perceptual measures. *American Journal of Speech-Language Pathology*, 28(3), 1115–1127. https://doi.org/10.1044/2019_AJSLP-18-0256

Wilson, S.M., Eriksson, D.K., Schneck, S.M. & Lucanie, J.M. (2018) A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLoS ONE*, 13(2), 1–29. https://doi.org/10.1371/journal.pone.0192773

Wilson, S., Roper, A., Marshall, J., Galliers, J., Devane, N., Booth, T. & Woolf, C. (2015) Codesign for people with aphasia through tangible design languages. *CoDesign*, 11(1), 21–34. https://doi.org/10.1080/15710882.2014.997744

Worrall, L., Sherratt, S., Rogers, P., Howe, T., Hersh, D., Ferguson, A. & Davidson, B. (2011) What people with aphasia want: their goals according to the ICF. *Aphasiology*, 25(3), 309–322. https://doi.org/10.1080/02687038.2010.508530

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

Zumbo, B.D. (2009) Validity as contextualized and pragmatic explanation, and its implications for validation practice. In *The concept of validity: revisions, new directions, and applications*, pp. 65–82. IAP Information Age Publishing.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Stark, B.C. & Dalton, S.G. (2024) A scoping review of transcription-less practices for analysis of aphasic discourse and implications for future research. *International Journal of Language & Communication Disorders*, 1–29. https://doi.org/10.1111/1460-6984.13028