
This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: http://openaccess.city.ac.uk/17105/

Link to published version: http://dx.doi.org/10.1111/1460-6984.12318

Copyright and reuse: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.
Reviewing the Quality of Discourse Information Measures in Aphasia

ABSTRACT

Background: Discourse is fundamental to everyday communication, and is an increasing focus of clinical assessment, intervention, and research. Aphasia can affect the information a speaker communicates in discourse. Little is known about the psychometrics of the tools for measuring information in discourse, which means it is unclear whether these measures are of sufficient quality to be used as clinical outcome measures or diagnostic tools.

Aims: The current review aimed to profile the measures used to describe information in aphasic discourse, and assess the quality of these measures against standard psychometric criteria.

Methods: A scoping review method was employed. Studies were identified using a systematic search of Scopus, Medline, and Embase databases. Standard psychometric criteria were used to evaluate the measures' psychometric properties.

Main contribution: The current review summarises and collates the information measures used to describe aphasic discourse, and evaluates their quality in terms of the psychometric properties of acceptability, reliability, and validity. Seventy-six studies described 58 discourse information measures, with a mean of 2.28 measures used per study (SD= 1.29, range 1-7). Measures were classified as functional measures (n= 33), which focused on discourse macrostructure, and functional and structural measures (n= 25), which focused on microlinguistic and macrostructural approaches to discourse as described by Armstrong (2000). There were no reports of the acceptability of data generated by the measures (distribution of scores, missing data). Test-retest reliability was reported for just 8/58 measures with 3/8 > 0.80. Intra-rater reliability was reported for 9/58 measures and in all cases % agreement was reported rather than reliability. Percent agreement was also frequently reported for inter-rater reliability, with only 4/76 studies reporting reliability statistics for 12/58 measures; this was generally high (> .80 for 11/12 measures). The majority of measures related clearly to the discourse production model described by Sherratt (2007), indicating content validity. 36/58 measures were used to make 41 comparisons.
between PWA and NHP, with 31/41 comparisons showing a difference between the groups. Four comparisons were made between genres, with two measures showing a difference between genres, and two measures showing no difference.

**Conclusions:** There is currently insufficient information available to justify the use of discourse information measures as sole diagnostic or outcome measurement tools. Yet the majority of measures are rooted in relevant theory, and there is emerging evidence regarding their psychometric properties. There is significant scope for further psychometric strengthening of discourse information measurement tools.
What we already know
Measuring aphasic discourse is a topic of increasing interest. Discourse can be measured in terms of the language a speaker uses, or functionally in terms of information a speaker can communicate. Although there is a wide range of discourse information measures available, very little is known about their quality in terms of psychometric properties.

What this paper adds
The current review summarises and synthesises the discourse information measures used to date; and it evaluates their psychometric properties of acceptability, reliability and validity.

What are the actual or potential clinical implications of this work
Very limited psychometric information is available about measures of discourse information. The field is still developing and further psychometric profiling of discourse information measures is essential before they can be adopted widely in clinical practice.
Introduction

Discourse can be defined as language beyond a single simple clause, used for a specific purpose (Armstrong, 2000; Halliday, 2004). Discourse is there the structure for much of the everyday communication that is fundamental to completing a range of daily activities (Davidson, Worrall, and Hickson, 2003). Everyday examples of discourse include giving instructions about how to carry out a procedure, such as assembling an item of flat pack furniture, or describing a beautiful view, or sharing a personal story. As discourse is an everyday use of communication, when aphasia affects the information conveyed in spoken discourse, there is a knock-on effect to a speaker’s activities and social participation (Davidson, Worrall, and Hickson, 2003).

Monologic discourse is typically elicited within aphasia treatment and research somewhat artificially, using probe questions and specific resources, such as picture description tasks (Linnik et al., 2015). Although this elicitation of monologic discourse may have limitations, such tasks represent a straightforward and easily constrained method of examining the multiplex of different discourse
types and genres likely to occur within spontaneous conversation (e.g., Boyle et al., 2011), whilst side-stepping the additional interactive processes which conversation entails (e.g., Beeke, 2012). There is evidence to suggest that speakers with aphasia produce monologic discourse impaired not only in terms of language functioning but also in terms of communicating information, and that the two impairments are not always correlated (see reviews by Armstrong, 2000; Linnik, Bastiaanse, and Höhlec, 2015, and Ellis et al., 2016). The current review is focussed on measures of information used to describe monologic spoken discourse in aphasia.

Spoken discourse can be unpicked at different levels, such as focusing on the words or structure of language that a speaker uses; and the information they communicate to their listener. Armstrong (2000) provides a useful framework for conceptualising these foci, and measures which reflect the different vantage points, including structuralist-orientated, functionalist-orientated, and those approaches which have aspects of both structural and functionalist orientations (figure 1). Measures which fall under the structural category seek to quantify how a speaker uses language in discourse; measures which fall under the functional category are those which focus on the transactional success of the discourse, in terms of 'the overall meaning of the text and the ways in which meanings are
organised within the text’ (Armstrong, 2000, p878). Between these categories, is a group of measures which both describe the transactional success of a speaker and also quantify the language used in discourse. Examples of the measures belonging to this ‘in between’ category include measures of how many main concepts a speaker communicates about a specific story, where main concepts represent a predetermined list of the key ideas in a discourse (Nicholas and Brookshire, 1995); and the number of Correct Information Units (CIUs) a speaker uses in their discourse, where a CIU is defined as a single word that is intelligible and relevant in context (Nicholas and Brookshire, 1993). Bryant et al (2016) described 565 of these structural measures in a recent systematic review. The current review will complement the review by Bryant et al., by focusing on those measures used to describe the ‘information’ level of communication, incorporating Armstrong’s categories of functional and structural and functional approaches; and by evaluating the quality of the measures terms of the psychometric properties of acceptability, reliability, and validity.

Communicating information, discourse, and aphasia

Authors’ final version of Pritchard, M., Hilari, K., Cocks. N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. *International Journal of Language and Communication Disorders.*
People with aphasia identify as a priority the ability to communicate information beyond basic needs and requirements, and the majority of examples they give of such communication reflect discourse targets (Worrall et al., 2011). Empirical research reflects this, with an increasing body of evidence suggesting that speakers with aphasia produce monologues which are impoverished in one or more ways in terms of information content (Linnik et al., 2015; Ellis et al., 2015). This finding presents a challenge due to the complex relationship between information content and language in discourse (Sherratt, 2007; Linnik et al., 2015). This complex relationship is likely to be heightened in speakers with aphasia, due to the nuances of aphasic language impairment at an individual level: whilst some speakers with aphasia have a relatively mild language impairment, they produce discourse containing limited information; whereas other speakers have a relatively severe impairment, but produce discourses containing a large amount of information.

There is a range of measures available to clinicians and researchers to describe the information speakers with aphasia include in discourse (see reviews by Armstrong, 2000, Linnik et al., 2015, and Ellis et al., 2016). However, the availability of measures is not enough to justify their use: in order to be used confidently in clinic and in research, measures need to be robust and of a high

Authors’ final version of Pritchard, M., Hilari, K., Cocks. N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. *International Journal of Language and Communication Disorders.*
quality. For example, they need to be grounded in relevant theory; measure a specific construct; and produce scores which are consistent across raters. These criteria reflect aspects of the *psychometric properties* of validity and reliability. If measures do not have strong psychometric properties (described in depth in Nunnally and Bernstein, 1994, and Streiner and Norman, 2000) they are unlikely to identify a problem, or to describe outcomes in a manner which inspires confidence. Currently, there is limited research describing the psychometric properties of discourse information measures, meaning their quality and whether they are appropriate for use is unclear.

The current review aimed to profile the information measures which have been used to describe discourse in speakers with aphasia, and evaluate their quality in terms of psychometric properties, using a scoping review method (Pham, Rajic, Greig, Sergeant, Papadopoulos and McEwen, 2014). The following section will outline these psychometric properties in more depth.

**Psychometrics**

Psychometrics is a field describing theory and technique around psychological measurement, and it provides frameworks for assessing the quality of measures. Standard techniques and criteria for evaluating the psychometric

Authors’ final version of Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. *International Journal of Language and Communication Disorders.*
properties of measures are outlined by Nunnally and Bernstein (1994), and Streiner and Norman (2000). The current review will focus on the properties of acceptability, reliability, and validity (including known-groups validity).

Acceptability looks at the quality of the data produced by a measure. This includes whether a measure yields a large proportion of missing data, and how its scores are distributed. In relation to discourse, missing data might mean that a participant was unable to retell any part of a Cinderella story (and so their discourse was not included in the subsequent analyses); or it might mean that a participant was unable to include in their discourse information about one picture in a sequence (with their descriptions of the other pictures included as a full discourse). Therefore, if data generated by using a measure contains a lot of missing values, it has poor acceptability. Score distribution relates to the assumption that any normal random variable will produce data that is symmetrically distributed in a bell curve. In relation to discourse, this can be considered in terms of normal variation when speakers tell stories: one would not necessarily expect all neurologically healthy speakers to produce a ‘textbook’ discourse on all occasions, and therefore score at ceiling on all discourse information measures on all occasions. In lay terms, some speakers are excellent storytellers, some are poor storytellers, and the majority of speakers fall...

Authors’ final version of Pritchard, M., Hilari, K., Cocks. N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. International Journal of Language and Communication Disorders.
between these points, producing an average discourse. This variation would be expected to be reflected in scores on a discourse information measure. If a discourse information measure yields skewed findings, with an atypical number of speakers scoring very high or very low, the measure may be of poor quality and not reflect the normal range of performance.

Reliability refers to the consistency and stability of a measure. Specific types of reliability include inter-rater reliability, intra-rater reliability, test-retest reliability, and internal consistency. Internal consistency reflects whether all items in a measure work together to measure the same construct. Test-retest reliability suggests a measure will yield the same score when a participant is tested again and no change has occurred. Similarly, a measure will yield the same score if it is scored by different individuals/raters (inter-rater reliability), and by the same rater at a different time-point (intra-rater reliability) (Nunnally and Bernstein, 1994; Streiner and Norman, 2000).

Construct validity is the extent to which a measure captures the construct it claims to (Nunnally and Bernstein, 1994). One aspect of construct validity is convergent validity, which explores if a measure correlates with other related measures. There are currently no widely used assessments which diagnose a ‘discourse impairment’ in speakers with aphasia, and thus it is not feasible to find

Authors’ final version of Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. International Journal of Language and Communication Disorders.
related measures against which to evaluate the convergent construct validity of discourse measures, and therefore this was not assessed in the current review. McNeil (2007) looked at concurrent validity in discourse, comparing four measures (# CIUS, % CIUs, CIUs/ min, % main concepts) across different discourse elicitation procedures, however this does not represent convergent validity, as it does not compare one measure against another measure. Content validity is the extent to which a measure captures relevant and important information about the underlying construct (Streiner and Norman, 2000). Content validity has two main aspects: coverage or comprehensiveness, and relevance to the concept being considered. As no discourse measure seeks to be a comprehensive measure of all aspects of discourse, the content validity of the measures included in this review was evaluated for relevance against current theory. The framework of current theory used as a benchmark in the current review is the discourse production model outlined by Sherratt (2007), described further below, which incorporates detail from a number of widely used and validated models of discourse and language production.

A further aspect of construct validity is known-groups validity. Known-groups validity evaluates whether a measure is able to discriminate known similarities and differences between groups (Nunnally and Bernstein, 1994).
the context of discourse measures in the current review, known groups validity will describe if a measure differentiates speaker group (participants with aphasia (PWA) vs. neurologically healthy participants (NHP)); and between different discourse genres (Boyle, 2011). Such information indicates whether a measure can discriminate sufficiently to be clinically useful.

**Theories of discourse production and structure**

In order to ascertain whether each measure reviewed has content validity, the current review explored if each measure is grounded within the discourse production theory described by Sherratt (2007). Sherratt’s model was chosen because it incorporates both cognitive and linguistic processing in a multilevel model, which is necessary because discourse production is complex and multifaceted. Sherratt’s model of discourse production builds on work completed by Frederiksen et al. (1990) on discourse comprehension. It describes the stages of discourse production from the initial trigger, through to articulation and, in doing so, it provides a framework that can be used for identifying which stage of processing each discourse measure addresses, and it forms the conceptual basis for hypotheses regarding underlying impairment.

In Sherratt’s model discourse starts as an idea which must be packaged for spoken language though a series of stages. Although these stages are

Authors’ final version of Pritchard, M., Hilari, K., Cocks. N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. *International Journal of Language and Communication Disorders.*
presented sequentially using boxes and arrows, it is likely that discourse 
production is an online and dynamic process, with some stages taking place 
repeatedly and/or simultaneously. This dynamism is reflected to some degree in 
the model, represented by the double-headed arrows, where the speaker 
monitors their discourse while they speak, and reconfigures it as appropriate.

The first step in Sherratt’s model, the *input trigger*, is where a speaker 
identifies a discourse they wish to communicate, for example, that they want to 
tell a specific story. In the next step, *frame/schema generation*, the speaker 
identifies an appropriate frame for their discourse, drawn from previously 
experienced discourse templates, which stipulate the information that ‘should’ be present, and the shape the discourse ‘should’ take. Next, the speaker moves on to the stages of *insertion and integration of information*, where semantic and episodic memory stores are accessed, and the information is added to the discourse frame. For example, in a procedural narrative, this could be the tools required to complete the procedure. Following this, *selection and topicalisation of information* takes place, and information relating to the discourse is sequenced and edited, based on a speaker’s pragmatic judgments. For example, in a procedural discourse about how to change the wheel on a car, if a speaker was advising someone who had never completed the procedure, the speaker would...
be likely to give more information than they would give to someone who had completed the procedure before. The next level, *generation selection and chunking of propositions*, is where logical relationships between information are assigned (such as foreground/background information, temporal sequences, or causation and consequence) and propositions are divided accordingly; these propositions and the relationships between them are then linguistically encoded at the next stage, *linguistic formulation*. Aphasia may present challenges to coherent production of an informative discourse, due to impairments relating to language itself, or due to challenges with preparing information for language (often referred to as thinking-for-speaking, see Dipper et al, 2005, for a review).

A clear strength of Sherratt’s model is its firm roots in well-known and widely accepted theories of the different aspects involved in discourse production. The model adds value to these theories by linking them together explicitly, and describing the stages from the trigger through to production. For example, ‘frame/ schema generation’ is in line with previously described theoretical constructs of *conceptual and suprastructural* frames (Fayol and Lemaire, 1993); the *macrostructure* of discourse (Kintsch and van Dijk, 1978); discourse frames (Frederiksen et al, 1990); and sentence production (Levelt,
1989). In linking these established theories, Sherratt’s model provides a useful basis for assessing the content validity of discourse information measures.

**The current review**

In the current review, discourse information measures will be described and grouped into categories based on whether they are *functional*, or describe aspects of both *function and structure*, as described by Armstrong (2000). The psychometric properties of acceptability, reliability and validity of the measures will be assessed against standard psychometric criteria. The review will aim to answer the questions:

1. What measures have been used to describe the information in aphasic discourse?
2. What is the quality of discourse information measures for speakers with aphasia, based on their psychometric properties (Nunnally and Bernstein, 1994; Streiner and Norman, 2000)?

**Methods**

A scoping review methodology was used for the current study (Pham et al., 2014; Arksey and O’Malley, 2005). This methodology shares features with a systematic
review, using rigorous and transparent methods to identify all the relevant literature relating to a research question. It differs from a systematic review, in that the primary purpose of the review is to map the current body of literature.

The process for searching, inclusion, and exclusion within the current review is summarised in the flowchart in figure 2. The databases searched were Scopus, Medline, and EmBase, using the terms “(‘discourse’ or ‘narrative’) and (‘analysis’ or ‘assessment’) and (‘aphasia’ or ‘dysphasia’)”. Records were screened using title and abstract, with the inclusion criteria of

- written in English;
- involving people with aphasia following stroke;
- reporting monologic discourse data in any language;
- describing measures which focus on information in discourse.

Whether or not measures focused on information in discourse was judged in line with the categories by Armstrong (2000), outlined in the Introduction section. Measures which were structural measures only were excluded.

It was anticipated that the search would yield measures used to analyse a range of different monologic discourses ed using a range of methodologies, in line with recent reviews of discourse production in aphasia (Bryant et al., 2016; Linnik et al., 2015). It was also anticipated that the search would yield descriptive
studies, and therapy studies, which used the measures but did not seek to profile their quality. Therefore, the evaluations in the current review are on the quality of the discourse measures, and by no means reflect the overall quality of the studies.

In order to answer RQ1, a short description of each measure used in the studies was extracted from the papers. Measures were viewed as a pool, and categories from Armstrong (2000) were used to identify measures as functional or functional and structural. Following this, a content analysis approach (Patton, 2002) was used to identify the kinds of information described by the measures. A second rater, a practising Speech and Language Therapist also coded the measures. Coding agreement between the raters was 100%.

In order to answer RQ2, measures were assessed against established criteria on psychometric properties (Nunnally and Bernstein, 1994; Streiner and Norman, 2000) (see table 1). Judgements regarding discourse genre were made using the discourse genres outlined by Boyle (2011, p1310), where narrative discourse is a description of events, procedural discourse provides directions or instructions, descriptive discourse describes something in detail, and expository discourse explains something in detail.

-------------------------------------------------------------------------Table 1 about here-------------------------------------------------------------------------

Authors’ final version of Pritchard, M., Hilari, K., Cocks. N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. International Journal of Language and Communication Disorders.
Results

Two hundred and eighty-five records were identified (n=273 through the database searching, n=11 through hand-searching reference lists, and n=1 from discussion with experts) (figure 2). After duplicates were removed, 257 records remained, which were screened for eligibility.

One hundred and seventy-three records were excluded as they did not meet the inclusion criteria for the review. The full texts of the remaining 84 papers were screened for eligibility and a further eight records were excluded. The remaining 76 papers were included in the review. These comprised 27 therapy studies, and 48 descriptive studies.

1. What measures have been used to describe the information present in aphasic discourse?

The 76 studies included 174 incidents of discourse measure use, which reflected 58 different measures of information (table 2). Thirty-three of the 58 measures were classified as functional measures, including measures of story grammar, utterance/propositional level information (including global coherence, local coherence, relevance, utterances with new information, categories of information, propositions, and violations); topic use, and overall listener judgment. The
remaining 25 measures were classified as structural and functional, including measures of single word/phrase level information (n=17), including correct information units, information units, content units, and lexical information units; and main concepts measures (n=8). Further descriptions about each group of measures are outlined in table 2.

2. What is the quality of discourse information measures for speakers with aphasia, based on their psychometric properties (Nunnally and Bernstein, 1994; Streiner and Norman, 2000)?

Largely, psychometric properties were not outlined for the measures reviewed in the current paper.

Acceptability

No data was given on the acceptability of the measures used, in terms of reporting any missing data and distribution of scores.

Reliability
Four studies described the test-retest reliability of eight measures using correlations (Boyle, 2014; Capilouto and Wright, 2006; Nicholas and Brookshire, 1994; Nicholas and Brookshire, 1995)\(^1\). Overall, three measures had a high level of reported test-retest reliability (>0.80) (single word/phrase level information (CIUs/min, accurate complete main concepts, and absent main concepts), and three measures which did not meet the threshold for test-retest reliability (# CIUs, accurate incomplete main concepts, and inaccurate incomplete main concepts). Percent CIUs and % main events measures both had equivocal findings. For the % CIUs measure, one study reported test-retest reliability above the threshold (Nicholas & Brookshire, 1994), whilst one study reported test–retest reliability below the threshold (Boyle, 2014, \(r=0.61\)). For the % main events measure, a single study reported different levels of test-retest reliability for people with and without aphasia, with test-retest reliability being over the threshold for PWA, and under the threshold for NHP (Capilouto & Wright, 2006, \(r=0.71\))

\(^1\) A further two studies explored this using percentage agreement statistics (Yorkston and Beukelman, 1980), and ANOVA statistics to describe test-retest differences (Cameron et al., 2010). These were excluded because the methods for reporting reliability because these did not meet the criteria for evaluating psychometric quality used in the current review (table 1).
Inter-rater reliability data is summarised in table 3. Overall, inter-rater reliability data were not reported for 59/174 incidences of measure use, with each measure subcategory containing some unreported data (percentage of unreported reliability data by measure subcategory: 

- **topic use** = 100%;
- **overall listener judgement** = 50%;
- **story grammar** = 61.11%;
- **utterance/propositional level information** = 52.38%;
- **main concepts** = 40.74%;
- **single word/phrase level information** = 17.7%.

The majority of reports of reliability used percentage agreement (102/174), which does not meet standard psychometric criteria for reliability checking (Hallgren, 2012). Just 6/76 studies, reporting on 14/58 measures (with each measure used once each) gave reliability data calculated using reliability statistics (Cronbach’s alpha, ICC, or kappa statistics). Altman (2014) reported reliability for three measures of **single word/phrase level information** and **utterance/propositional level information** (# CIUs, # utterances with new information, # propositions) using Cronbach’s alpha, with high levels of agreement at >0.91 and >0.98 respectively. Andreetta et al. (2012) reported kappa >0.80 for four measures of **utterance/propositional level information** and **single word/phrase level information** (% global coherence errors, % local coherence errors, # main concepts, % lexical information units). Finally, ICCs
were used in four studies (Hula, 2003; Kendall, 2008, Ulatowska et al., 1983a, Ulatowska et al, 1983b), reporting reliability of single word/ phrase level information measures (% direct information units, % alternative information units, # total information units, information units/ minute, a measure of CIUs, and two measures of overall listener judgements), and two measures of overall listener judgements. Five of these seven measures met the threshold for ICC > .80, and two measures fell below this threshold: the measures of CIUs (r= 0.70, Kendall, 2008) and one overall listener judgements measure used on narrative discourse (r range 0.79-0.91, Ulatowska et al, 1983).

Ten studies reported intra-rater reliability, using 9/58 measures of single word/ phrase level information, function of propositional information, and main concepts (% CIUS n=7; CIUS/ min n=7; # CIUs n=4; global coherence n=1; local coherence n=1; accurate complete main concepts n=1; accurate incomplete main concepts n=1; inaccurate main concepts n=1; absent main concepts n=1). In all 24 instances of the measures' intra-rater reliability being reported, percent agreement was given, which was >80%.

-----------------------------------table 3 about here-----------------------------------

Content validity (theory)
The majority of measures related to the frame/schema generation level of Sherratt’s model, as they describe the information present or absent, or the hierarchical or causal relationships between that information. Measures from the functional and structural and functional category relate to this level: story grammar (n=9), and main concepts (n=8). Story grammar measures reflect the assumption that a discourse frame should include different ‘kinds’ of information, such as ‘orientation’ and ‘initiating event’. Main concepts measures describe the same discourse frame in a different way. They identify specific information that ‘should’ be present in a specific discourse by analysing the information that a group of neurologically healthy speakers include in a discourse.

The category of function of utterance or proposition measures (n=13) are likely to reflect the level of selection and topicalisation of information, where information in the discourse frame interacts with pragmatic judgements, and speakers identify the information to include and exclude. Topic use measures (n=3) reflect the next stage of generation and chunking of information, where speakers organise the information and relationships between it. The single word/phrase level information measures (n=17) reflect multiple levels as these measures are structural and functional, and reflect both word production and aspects such as intelligibility and relevance. Therefore, these measures are likely
to reflect the levels of *articulation, linguistic formulation, and selection and topicalisation of information.*

The *functional* measures which linked less clearly to Sherratt’s model were *overall listener judgements* (n=8). These measures used judgements from naïve listeners (Behrs et al., 2009; Cupit, Rochon, Leonard, and Laird, 2010; Jacobs, 2001); and expert listeners (Cupit et al., 2010, Ulatowska, 2003; Ulatowska, 2001; Ulatowska et al, 2013, Ulatowska et al., 1983; Ulatowska et al., 1983). Rather than relating to specific stages of production or the structure of discourse, this group of measures instead focus on a listener’s perception of the discourse overall.

**Construct validity (Known groups validity)**

Across the 76 studies reviewed, 41 statistical comparisons were made between participant groups (between the discourses produced by NHP and those produced by PWA - table 4), using 36 measures (*single word/ phrase level information* n=11; *utterance/ propositional level information* n=8; *main concepts* n=7; *overall listener judgement* n= 4; *story grammar* n= 4; *topic use* n=2). The majority of comparisons (31/41) indicated a difference between the groups, with all comparisons for the measures of *topic use* (n=2) and *overall listener judgement* (n=4) indicating a difference between the groups.
The majority of studies reported data on more than one discourse, and often reported discourse data related to more than one discourse genre. However, only two comparisons were made between discourse genres (Altman et al., 2014, Capilouto et al. 2006), using a total of four measures (single word/phrase level information (n=1), utterance/propositional level information (n=2), and main concepts (n=1)). The single word/phrase level information measure, #CIUs, and the main concepts measure, % main concepts, showed a difference between the genres, whilst the utterance/propositional level information measures, # utterances with new information, and # utterances with interpretable meaning, did not show a difference between the genres.

Summary
The search yielded 76 studies which, as expected, covered a range of discourse elicitation methods, and covered both descriptive studies and therapy studies which did not seek to profile the measures. These 76 studies contained 174 incidents of discourse information being measured, using 58 discourse information measures which were either functional measures or structural and...
functional measures (Armstrong, 2000). The functional categories included subcategories of story grammar, utterance or propositional level information, topic use, and listener judgement. The function and structural category included measures of single word/phrase level information, and main concepts. Overall, certain measures of main concepts and single word information measures in the form of CIUs emerged from the current review as the reporting the strongest psychometric properties, as they met the thresholds for test-retest reliability, for interrater reliability, content validity, and reported data relating to known groups, in the form of participant groups and for genre. However, it is important to note that this strength is in the context of limited data regarding psychometric quality for the majority of measures reported in the current review.

Discussion

The current review summarised, described, and synthesised the discourse information measures used to analyse the discourse of people with aphasia. It also assessed these measures against standard psychometric criteria (Nunnally and Bernstein, 1994; Streiner and Norman, 2000). In doing so, the current review builds on previous reviews of discourse in aphasia (Armstrong, 2000; Bryant et al. 2016; Linnik et al, 2015). Overall, the review revealed that measures of main
concepts and single word information in the form of CIUs are emerging as the strongest measures profiled in the current review. However, there is limited information available regarding the psychometric quality of the majority of discourse measures in aphasia. Further information is required to strengthen the psychometric profiles of these measures in order that they can be used to diagnose impairment, or measure change in clients with aphasia.

There was a good deal of variability between measures which reflected the same or similar constructs (table 2). This heterogeneity included variability in the way in which constructs were framed (e.g., different approaches toward describing story grammar), and the ways in which these constructs were measured. Overall, such variability is likely to make synthesis of findings and approaches challenging. For example, in the category of story grammar, some measures described narrative discourse (Altman et al., 2012; Coelho et al., 1994; Li, 1995; Olness et al., 2010; Stark, 2010; Ulatowska, Freedman Stern, et al., 1983; Ulatowska et al., 1981; Ulatowska et al., 2004; Whitworth, 2010), and others described procedural discourse (Li, 1995; Ulatowska et al., 1981; Ulatowska et al., 1983; Weiss Doyel, et al., 1983). One story grammar measure collapsed data across genres (recount, expository, narrative, and procedural) and then
tallied the total number of information points a speaker communicated, such as orientation, body, and conclusion (Whitworth et al., 2015). Given the disparate nature of the measures, the framework by Sherratt (2007) provides an essential benchmark against which to evaluate whether each discourse measure captures relevant and important information about the underlying construct. When relating the measures to this model, the approaches to story grammar measurement are not as disparate as they might appear. Each story grammar measure has common roots in the narrative analysis framework outlined by Labov (1972), and reflects the same level of the model by Sherratt, suggesting each has content validity. Disparity between measures which share a common root construct is visible in other measure categories in the current review, such as the single word/phrase level information measures, despite each focusing on a common underlying construct. In future work, greater consistency in measurement approaches across studies would aid synthesis of findings, supporting clinicians’ and researchers’ interpretation of research findings.

Very limited psychometric information was available for the measures of discourse profiled, meaning that the majority of measures in the current review are questionable for use in diagnosis and outcome measurement. No studies
reported the acceptability of their data in terms of the distribution of scores, or missing data. This lack of information means that: there is no evidence that the measurement tools are acceptable to PWA; there is no clear idea of what ‘normal’ performance looks like on the measures profiled; nor whether the data from these measures should be analysed using parametric statistical tests, which rely on data being normally distributed. This is particularly important, given the likely variability inherent in discourse production: it is unlikely that neurologically healthy speakers behave in a singular manner when producing discourse. Until further information is available on normal performance and variation in discourse production, identifying a discourse impairment arising from aphasia is likely to represent an ongoing challenge. Furthermore, studies comparing groups using parametric statistics need to be interpreted with caution until further information regarding score distribution of measures is available.

Although reliability data in the current review was limited, with the majority of measures not reporting test-retest reliability and inter-rater reliability, where realibility statistics were reported, they they were high. For example some of the measures of CIUs and Main Concepts achieved both high test-rest reliability ratings(CIUs per min; # accurate and incomplete Main Concepts) and high
inter-rater reliability statistics (\# CIUs, \#Main Concepts). This review finding suggests that information in discourse can be measured reliably between different raters, and different time-points. This is an encouraging finding, as many of the information measures outlined within the current review appear subjective. Future work should aim to report reliability using ICCs or kappa statistics as appropriate because it may well be the case that other measures are as reliable. Furthermore, future work should report reliability data clearly for each measure and should distinguish between different ways of measuring a single construct (e.g. \# vs \%). In a number of studies in the current review, reliability data was pooled across measures of the same or differing categories, meaning the reliability of each measure was unclear. For example, a number of studies reported a single reliability figure relating to the identification of CIUs, when they had used multiple CIU measures (% CIUs, CIUs/ min, etc).

The majority of measures in the current review were based on the underlying theories and could be related to the model of Sherratt (2007), and thus had content validity. This is positive for future clinical application and use because if a measure can be directly related to a model of how a speaker communicates information in discourse, then clinicians can more confidently
assume that a key element of discourse processing has been measured, giving a firm basis for diagnosis and treatment planning. Only the measures of overall listener judgment did not relate clearly to Sherratt’s theory, although these measures are likely to reflect theoretical underpinnings of discourse indirectly. For example, a measure focusing on a listener’s judgment of the overall ‘coherence’ of the whole discourse may reflect the level of frame/schema generation on Sherratt’s model, but the measure does so by seeking a listener’s judgement of the acceptability of the discourse frame, rather than seeking to measure the discourse frame directly. The overall listener judgement measures are likely to be ‘ecologically valid’, with listener judgements linked to linguistic aspects of a speaker’s discourse (e.g. Jacobs, 2001).

When categorising the measures into the categories of functional and functional and structural in the current review (Armstrong, 2000), we encountered a challenge in that we identified examples of functional measures which appeared to conflated structure and function, despite the fact that the primary aim of the measure was functional. For example, the story grammar measure described by Ulatowska et al. (1983) (which is included as an example of a functional approach to measurement, given by Armstrong, 2000), quantifies the information components in a discourse by counting the number of clauses a
speaker uses in each information component: ‘in comparing the performance of aphasic and control subjects in preserving narrative structure, clauses were classified into four categories: setting, action, resolution and evaluation’ (Ulatowska et al., 1983, p 325). That is to say, if clauses are determined using grammatical criteria (rather than, for example, representing utterances, which are often not determined using syntactic criteria), a speaker’s scores using this functional measure may be compromised, meaning it may be a structural and functional measure. Similar ambiguity can also be observed in a number of the measures of utterance/propositional information, global and local coherence. There is, therefore, a need for clearer distinctions regarding what a measure is reflecting, to avoid inadvertently penalizing speakers, or confounding findings; and for work to give examples to illustrate how the measure is used, to make these distinctions as clear as possible.

The majority of comparisons between PWA and NHP found a difference, suggesting that functional and functional and structural measures have known-groups validity, as they differentiate between the groups. However, a number of the sub-categories had equivocal findings. In these cases, the lack of findings indicating group difference may be due to variability in the measures, and metrics used to describe them (e.g., % CIUs, # CIUs, CIUS/minute). However due to

Authors’ final version of Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. International Journal of Language and Communication Disorders.
the extent of heterogeneity in measures, it is beyond the scope of the current paper to go into the impact of this variability in depth. On the other hand, the lack of a group difference for some of the sub-categories of measure may reflect a true lack of difference between the groups for a specific construct or sub-construct. Interesting to note is the fact that the category of *overall listener judgement* (n=8), a group of measures likely to be ecologically valid, represents one of just two categories of measure consistently showing a difference between neurologically healthy speakers and speakers with aphasia. There is significant scope for investigating which of the measures in the current review relate most clearly to these ecologically valid judgements of clinicians, researchers, naïve listeners, and people with aphasia themselves (such as the indicative findings of Jacobs 2001, linking % CIUs with listener judgements).

Difference between genres is described within the theoretical framework of discourse production (Sherratt, 2007), and throughout the literature on discourse production in aphasia (e.g., Whitworth et al., 2015; Linnik et al, 2015). Although there were a broad range of discourse genres reflected in the studies in the current review, including genres of narrative, procedural, descriptive, and expository discourse, many studies aggregated data across discourse genres,

---

Authors’ final version of Pritchard, M., Hilari, K., Cocks. N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. *International Journal of Language and Communication Disorders.*
and only two studies (measures n=4) compared discourse genres. More work is required before firm conclusions can be drawn regarding these measures. Furthermore, more information is required about how discourse varies across genre, in order to ascertain what normal variation is between genres and what reflects impairment; and in order to understand variability of speakers’ discourse profiles. In order to do this, there is a need for measures which are flexible, and can be used across different discourse genres. Comparing across discourse genres, and adapting measures in order that they can achieve such a comparison would be a valuable direction for future work.

**Implications and future research**

The current review identified a broad range of discourse information measures available, but the evaluation of the measures’ psychometric properties revealed that the majority do not have sufficiently strong acceptability, reliability and validity to justify their use as outcome measures or diagnostic tools. Discourse is described as a priority for therapy for people with aphasia (Worrall et al., 2011) and is increasingly the subject of assessment and therapy research in speakers with aphasia (see review by Bryant et al., 2016). However, without an evaluation of the strength of the psychometrics underpinning discourse
measures, it is unclear if findings from this growing body of research are reliable or are simply a result of poor quality measures (such as measures with skewed distribution). It is therefore of fundamental importance that such information is available, and it should represent a priority for future research. Future research should aim to profile how measures discriminate between known groups; their stability over time; if measures are ecologically valid, reflecting the views of speakers with aphasia, clinicians and researchers; and whether these measures can be used clinically.

**Conclusions**

The current review profiled the measures for assessing information in discourse in speakers with aphasia, and assessed their quality against standard psychometric criteria. Although there are a range of measures which have been used to assess information discourse by speakers with aphasia, more information is needed to justify their use as tools in diagnosis or treatment. There is a promising foundation, in that the majority of measures relate clearly to a theoretical model of discourse production; and emerging reliability and known groups validity data is positive. From the evidence currently available, certain measures of CIUs (i.e., #CIUs but not %CIUs) and main concepts (again # rather
than %) emerge as the most reliable; and # CIUs and % main concepts (rather than #main concepts in this case) having the strongest known groups validity.

Future work is needed to consolidate findings, and to validate the measures further, leading to increased confidence in the use of discourse information measures in research and clinical practice.

REFERENCES
ANTONUCCI, S., 2009, Use of semantic feature analysis in group aphasia treatment. Aphasiology, 23 (7-8), 854-866

Authors’ final version of Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. International Journal of Language and Communication Disorders.


MARSHALL, R., LAURENS-GORE, J., DUBAY, M., WILLIAMS, T., and BRYANT, D., 2015, Unilateral forced nostril breathing and aphasia: Exploring unilateral forced nostril breathing as an adjunct to aphasia treatment a case series. Journal of Alternative and Complimentary Medicine, 21 (2), 91-99


MCNEIL, M., DOYLE, P., PARK, G., FOSSETT, T., and BRODSKY, M., 2002, Increasing the sensitivity of the story retell procedure for the story retelling procedure. Aphasiology, 16 (8), 815-822

Authors’ final version of Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. International Journal of Language and Communication Disorders.

Figure 1: Approaches to discourse measurement from Armstrong (2000)
Discourse sample

Pragmatic judgments and requirements

Semantic memory

Episodic memory

Input trigger

Frame/ schema generation

Insertion of semantic information

Integration of semantic information

Selection and topicalisation of information

Generation selection and chunking of propositions

Linguistic formulation

Articulation

Top down and bottom up processing

Figure 2. Discourse production model from Sherratt (2007)
Authors’ final version of Pritchard, M., Hilari, K., Cocks. N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. *International Journal of Language and Communication Disorders.*
Records identified through database searching (n = 273)

Additional records identified through other sources (n = 12)

Records after duplicates removed (n = 257)

Records screened (n = 257)

Records excluded (n = 173)

Full-text articles assessed for eligibility (n = 84)

Full-text articles excluded (n = 8)

Articles (n = 76)

Authors’ final version of Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. Reviewing the Quality of Discourse Information Measures in Aphasia. *International Journal of Language and Communication Disorders.*
Figure 3: Flowchart of studies included in the review
Table 1:
Definitions and criteria for psychometric properties (based on Nunnally and Bernstein, 1994; and Streiner and Norman, 2000)

<table>
<thead>
<tr>
<th>Psychometric property</th>
<th>Definitions</th>
<th>Quality criteria</th>
</tr>
</thead>
</table>
| 1. Acceptability      | The overall quality of the data, assessed by completeness of the data and score distribution | Missing data <10%  
Skewness between -1 and +1 |
| 2. Reliability        | Stability of a measuring instrument, assessed by administering the instrument to respondents on two different occasions (typically within 2-14 days) and examining the correlation between test and retest scores | Test-retest reliability: ICC > 0.75  
Inter-rater reliability: ICC > 0.80  
Intra-rater reliability: ICC > 0.80 |
| 3. Validity           | Evidence that a single entity is being measured, that scales are consistent with a conceptual model | Content validity: Evaluated qualitatively by checking relevance against theoretical model (Sherratt, 2007)  
Construct validity - Known groups: Significant differences in discourse scores between different groups (e.g. persons with aphasia vs. neurologically healthy persons). |
Table 2

Information measures in the current review (studies n= 76; measures n= 58, incidences of measure use= 174)

<table>
<thead>
<tr>
<th>Category (Armstrong, 2000)</th>
<th>Measure type (n= # measures)</th>
<th>Description</th>
<th>Studies (n= # measures in each study)</th>
<th>Measure (s) (n= incidences of measure)</th>
<th>Elicitation methods*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional (33)</td>
<td>Story grammar (n=9)</td>
<td>Specific information components and networks within a discourse</td>
<td>Altman et al. (2012); Capilouto and Wright (2009); Coelho et al. (1994); Hickin et al. (2015); Li (1995) (n=2); Cliness et al (2010); Purdy (2002); Stark (2010); Ulatowska et al. (1981) (n=3); Ulatowska et al. (1983a); Ulatowska et al. (1983b); Ulatowska et al. (2004); Ulatowska et al. (2013); Whitworth (2010); Whitworth et al. (2015).</td>
<td># points of information, e.g., sequences of time or causation; reference, and evaluation (n=6); % points of information (n=1) # utterances or clauses belonging to points of information e.g., orientation, coda (n=4) Presence/ absence of basic narrative structure (n=2) # temporo-causal sequences (n=1) # complete episodes in a story (initiating event+ action+ consequence) (n=1)</td>
<td>Narrative discourse, including Cinderella, retell and generation discourses; and personal discourse, including frightening experiences, recent vacations, company funded by the speaker, and participation in a sign language project Procedural discourse, including brushing teeth, combing hair, making sandwiches, changing a tyre Expository discourse, providing opinions on bullying, obesity, and global warming Descriptive discourse, using pictures</td>
</tr>
</tbody>
</table>

2 Figures are not reported for elicitation methods, as not all studies reported this clearly enough for comparison. A number of studies elicited multiple discourses, and gave only a broad summary of how these were elicited.
<table>
<thead>
<tr>
<th>Utterance/propositional level information (n=13)</th>
<th>Analysis of utterance/propositional level information analysis</th>
<th># points of information- procedural (e.g., steps, optional steps) (n=1); % points of information- procedural (n=1)</th>
<th>Point awarded for procedures where the proper sequence maintained, and enough detail provided that the task could be carried out (n=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance/propositional level information analysis</td>
<td>Altman et al. (2014) (n=2); Andreetta et al. (2012) (n=2); Christiansen (1995) (n=4); Christiansen (1999); Coelho and Flewellyn (2003) (n=2); Glosser and Deser (1990); Hickin et al. (2015); Marini et al. (2011) (n=2); O'Leary and Englebretson (2011); Hickin et al., (2015); Glosser and Deser (1990); Marini et al. (2011); Ulatowska et al. (2001); Ulatowska et al., (2004)</td>
<td>Local coherence, using a 5-point scale (n=3), scoring how well each utterance relates to the previous utterance # propositions (n=4). Utterances with interpretable meaning isolated and extracted. Global coherence scores using a 5-point scale, describing how well each utterance relates to the overall topic (n=2) % Global coherence errors (n=2) % Local coherence errors (n=2) Global coherence scores using a 4-point scale, describing how well each utterance relates to the overall topic (n=1) # Utterances that are coherent and relevant, whilst adding information a listener can identify as new (n=1) % propositions that are events vs. states vs. elaborations (n=1) % propositions that are story vs interaction focused (n=1) % propositions that are info gaps, repetitions, and irrelevant (n=1) Relevance, using a 7-point scale (n=1) Each piece of information rated, based on Narrative discourse, including Cinderella sorties, cartoon sequences, retelling stories e.g., ‘the bear and the fly’; personal discourses including narratives about family, work experience, experience from the past, and a frightening experience Expository discourse, including open ended questions Descriptive discourse, including picture descriptions (Norman Rockwell, ‘Picnic’ from the Western Aphasia Battery, and ‘Cookie Theft’ from the Boston Tests.</td>
<td></td>
</tr>
</tbody>
</table>
| Topic use  
| (n=3) | How a topic is divided into topics and smaller sub-topics within a discourse | Armstrong et al. (2007) (n=2); Armstrong (2011) (n=2) | # topics, subtopics, sub-subtopics, and sub-sub-subtopics (n=2) 

% topics, subtopics, sub-subtopics, and sub-sub-subtopics (n=1) 

# C-units per subtopic (n=1) | Narrative discourse, including personal stroke experience story 

_Procedural discourse_, including changing a lightbulb and making a cheese sandwich 

_Descriptive discourse_, including the 'Cookie Theft' from the Boston tests.

relevance to the characters and plot of the story. 

% propositions coded as on and off the main event line (n=1) 

# locations of direct expressions of fear (n=1)
| Overall listener judgement (n=8) | Ratings from expert or non-expert raters, scoring features of a whole discourse, such as 'coherence' and 'plot' | Behns et al. (2009); Cupit et al. (2010); Jacobs (2001); Ulatowska et al. (1983) (n=2); Ulatowska et al. (2003); Ulatowska et al. (2001); Ulatowska et al. (2013) | Scoring features such as 'difficult/easy to understand', effectiveness, listener comfort, coherence, and discourse quality using 2 point, 4 point, and 7 point scales, or using direct magnitude estimation | Narrative discourse, including Cinderella stories, cartoon sequences, fable retells, and personal stories including memorable experiences, and stories of 'I have never been so afraid'.

**Procedural discourses**, including changing a lightbulb, making a sandwich, making scrambled egg, and shopping in a supermarket

**Descriptive discourse**, including single picture description. |
<table>
<thead>
<tr>
<th>Structural and functional (measures= 25)</th>
<th>Single word/phrase level information (n=17)</th>
<th>Single word measures, where each single word is judged according to its relevance to the story</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albright and Purves (2008) (n=3); Altman et al. (2014); Andreotta et al., (2012); Antonucci (2009) (n=3); Ballard et al. (1999) (n=2); Boa and Rose (2011) (n=3); Boyle (2014) (n=3); Brennaise-Sarshand et al. (1991) (n=4); Brodsky et al. (2003); Brookshire and Nicholas (1994) (n=3); Cameron et al. (2006) (n=3); Cameron et al. (2010) (n=3); Cherney (2010a); Cherney (2010b); Cherney et al., (2012); Correia et al. (1990) (n=3); Craig (1993) (n=2); Doyle et al. (1995) (n=2); Doyle et al. (1998) (n=3); Doyle et al. (2000) (n=2); Edmonds and Babo (2011); Edmonds et al. (2014) (n=2); Fergadiotis et al. (2015); Fink et al., (2008) (n=2); Furnas (2003) (n=2); Georgeadis (2004); Gordon (2008); Hoover (2015); Hula (2003) (n=4); Jacobs (2001) (n=2); Kendall (2008); Knoph et al. (2015) (n=2); Linebarger (2007) (n=2); Marshall et al. (2015); McNeil (2001) (n=4); McNeil (2002) (n=2); McNeil (1997); McNeil (2007) (n=4); Murray (2004); Murray (2007) (n=3); Murray (1998) (n=2); Murray (2000); Nicholas and Brookshire (1993) (n=2); Ross, 1999 (n=2); Stark, 1999 (n=2); Ulatowska et al. (2004)</td>
<td>Albright and Purves (2008) (n=3); Altman et al. (2014); Andreotta et al., (2012); Antonucci (2009) (n=3); Ballard et al. (1999) (n=2); Boa and Rose (2011) (n=3); Boyle (2014) (n=3); Brennaise-Sarshand et al. (1991) (n=4); Brodsky et al. (2003); Brookshire and Nicholas (1994) (n=3); Cameron et al. (2006) (n=3); Cameron et al. (2010) (n=3); Cherney (2010a); Cherney (2010b); Cherney et al., (2012); Correia et al. (1990) (n=3); Craig (1993) (n=2); Doyle et al. (1995) (n=2); Doyle et al. (1998) (n=3); Doyle et al. (2000) (n=2); Edmonds and Babo (2011); Edmonds et al. (2014) (n=2); Fergadiotis et al. (2015); Fink et al., (2008) (n=2); Furnas (2003) (n=2); Georgeadis (2004); Gordon (2008); Hoover (2015); Hula (2003) (n=4); Jacobs (2001) (n=2); Kendall (2008); Knoph et al. (2015) (n=2); Linebarger (2007) (n=2); Marshall et al. (2015); McNeil (2001) (n=4); McNeil (2002) (n=2); McNeil (1997); McNeil (2007) (n=4); Murray (2004); Murray (2007) (n=3); Murray (1998) (n=2); Murray (2000); Nicholas and Brookshire (1993) (n=2); Ross, 1999 (n=2); Stark, 1999 (n=2); Ulatowska et al. (2004)</td>
<td>Albright and Purves (2008) (n=3); Altman et al. (2014); Andreotta et al., (2012); Antonucci (2009) (n=3); Ballard et al. (1999) (n=2); Boa and Rose (2011) (n=3); Boyle (2014) (n=3); Brennaise-Sarshand et al. (1991) (n=4); Brodsky et al. (2003); Brookshire and Nicholas (1994) (n=3); Cameron et al. (2006) (n=3); Cameron et al. (2010) (n=3); Cherney (2010a); Cherney (2010b); Cherney et al., (2012); Correia et al. (1990) (n=3); Craig (1993) (n=2); Doyle et al. (1995) (n=2); Doyle et al. (1998) (n=3); Doyle et al. (2000) (n=2); Edmonds and Babo (2011); Edmonds et al. (2014) (n=2); Fergadiotis et al. (2015); Fink et al., (2008) (n=2); Furnas (2003) (n=2); Georgeadis (2004); Gordon (2008); Hoover (2015); Hula (2003) (n=4); Jacobs (2001) (n=2); Kendall (2008); Knoph et al. (2015) (n=2); Linebarger (2007) (n=2); Marshall et al. (2015); McNeil (2001) (n=4); McNeil (2002) (n=2); McNeil (1997); McNeil (2007) (n=4); Murray (2004); Murray (2007) (n=3); Murray (1998) (n=2); Murray (2000); Nicholas and Brookshire (1993) (n=2); Ross, 1999 (n=2); Stark, 1999 (n=2); Ulatowska et al. (2004)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Main concepts (n= 8)</th>
<th>Inclusion of predetermined main events or concepts in a discourse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albright and Purves (2008); Andreotta et al. (2012); Armstrong et al (2007) (n=4); Boyle (2014) (n=4); Capilouto et al. (2006); Cupit et al. (2010), Doyle et al. (1995); Doyle et al. (1998); Doyle (2000); Gleason et al. (1980); Marini et al., (2011); McNeil (2001); McNeil (2007); Nicholas and Brookshire (1995) (n=4); Ross (1999); Stark (2010); Ulatowska et al.,</td>
<td>Albright and Purves (2008); Andreotta et al. (2012); Armstrong et al (2007) (n=4); Boyle (2014) (n=4); Capilouto et al. (2006); Cupit et al. (2010), Doyle et al. (1995); Doyle et al. (1998); Doyle (2000); Gleason et al. (1980); Marini et al., (2011); McNeil (2001); McNeil (2007); Nicholas and Brookshire (1995) (n=4); Ross (1999); Stark (2010); Ulatowska et al.,</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% of story propositions, thematic units, main concepts or main events (n= 5)</th>
<th>% core propositions, key propositions, and main themes (n=4)</th>
<th>% Accurate complete main concepts (n=4)</th>
<th>% Accurate incomplete main concepts (n=4)</th>
<th>% Incomplete main concepts (n=4)</th>
<th>% Absent main concepts (n=4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narrative discourse, including Cinderella, Red Riding Hood, cartoon sequence storytelling, ‘Bear and Hippo; personal narratives, including a tip or a happy memory, and ‘tell me what you usually do on a Sunday’</td>
<td>Descriptive discourse, including picture descriptions, such as the Norman Rockwell pictures, the Cookie Theft</td>
<td>Procedural discourse, including making a sandwich, and making scrambled eggs</td>
<td>Expository discourse, including a description of post-stroke speech</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abbreviations: # = number, %= percent, /= per, CIUs= Correct Information Units, single words, intelligible and relevant in context, IUs= Information Units, single words produced in a specific discourse are compared with specific words produced by a control group, CUs= Content Units, single words produced in a specific discourse are compared with specific words produced by a control group, LIUs= Lexical Information Units, defined as words that are phonologically well formed, and appropriate from a grammatical and pragmatic perspective.
Table 3

Inter-rater reliability data for measures (measures n= 58, total incidence of measure use n= 174)

<table>
<thead>
<tr>
<th>Category (Armstrong, 2000)</th>
<th>Measure category</th>
<th>Total frequency of measure</th>
<th>Not reported (NR)</th>
<th>Kappa &gt;0.80</th>
<th>Cronbach's alpha &gt;0.80</th>
<th>ICC</th>
<th>Percentage agreementa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional (50)</td>
<td>Story grammar</td>
<td>18</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Function of Utterance/propositional level</td>
<td>21</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Topic Use</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Overall listener judgement</td>
<td>8</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

3 Percentage agreements were not calculated in a uniform manner across studies. For example, some studies reported an overall agreement percentage, whilst others calculated agreement per participant/ subgroup/ discourse/ measure, and then calculated an agreement mean and range. This category therefore includes multiple calculation methods. Where mean and range was included, mean agreement was used for categorisation. Where mean was not included, single agreement figure, or the lower limit of the agreement range was used for categorisation.

4 Figure is not ICC but the estimated reliability of the pooled (mean) rating for three raters, using a Spearman-Brown Reliability correlation
<table>
<thead>
<tr>
<th>Functional and structural (n=105)</th>
<th>Single word</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>96</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1 (0.70)</td>
<td>71</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main concepts</td>
<td>27</td>
<td>11</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>174</td>
<td>58</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>99</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 4

**Statistical comparisons between PWA and NHP in the current review (n=41)**

<table>
<thead>
<tr>
<th>Analysis</th>
<th>NHP ≠ PWA (n=31)</th>
<th>NHP = PWA (n=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance/propositional level information</td>
<td>Andreetta et al. (2012); Christiansen (1995) (n=4); Ulatowska et al. (2001)</td>
<td>Andreetta et al. (2012); Glosser and Deser (1990) (n=2); Olness and Englebretson (2011);</td>
</tr>
<tr>
<td>Topic use</td>
<td>Armstrong (2011) (n=2)</td>
<td></td>
</tr>
<tr>
<td>Overall listener judgement</td>
<td>Behrens et al. (2009); Ulatowska, Freedman Stern, et al. (1983); Ulatowska, Weiss Doyel, et al. (1983); Ulatowska et al. (2001)</td>
<td></td>
</tr>
<tr>
<td>Main concepts</td>
<td>Capilouto et al. (2006); Nicholas and Brookshire (1995) (n=4); Ulatowska, Freedman Stern, et al. (1983)</td>
<td>Andreetta et al. (2012)</td>
</tr>
<tr>
<td>Single word/phrase level information</td>
<td>Andreetta et al. (2012); Breennaise-Sarshand et al. (1991) (n=3); Brodsky et al (2003); Correia et al (1990); McNeil (2001); McNeil (2002); Nicholas and Brookshire (1993) (n=2); Yorkston and Beukelman (1980)</td>
<td>Breennaise-Sarshand et al. (1991); Correia et al (1990) (n=2)</td>
</tr>
</tbody>
</table>