Abstract

Background. Discourse in adults with aphasia is increasingly the focus of assessment and therapy research. A broad range of measures is available to describe discourse, but very limited information is available on their psychometric properties. As a result, the quality of these measures is unknown, and there is very little evidence to motivate the choice of one measure over another.

Aims. The current study explored the quality of a range of discourse measures, targeting sentence structure, coherence, story structure and cohesion. Quality was evaluated in terms of the psychometric properties of acceptability (data completeness and skewness), reliability (inter- and intra- rater), and validity (content, convergent, discriminant, and known groups).

Methods and Procedures. Participants with chronic mild to moderate aphasia were recruited from community groups. Participants produced a range of discourses which were grouped into Cinderella and Everyday discourses. Discourses were then transcribed orthographically, and analysed using four macrolinguistic and one microlinguistic measure (Story Grammar, Topic Coherence, Local Coherence, Reference Chains, and Predicate Argument Structure). Data were evaluated against standard predetermined criteria to ascertain the psychometric quality of the measures.

Outcomes and Results. Seventeen participants took part. All measures had high levels of acceptability, inter- and intra- rater reliability, and had good content validity, as they could be related to a level of the theoretical model of discourse production. For convergent validity, as expected, 8/10 measures correlated with the WAB-R spontaneous speech scores, and 7/10 measures correlated with the Kissing and Dancing Test scores ($r>0.3$) giving an overall positive rating for construct validity..). For discriminant validity, as expected, all measures had low correlations with RCPM and WAB-R Auditory Verbal Comprehension scores ($r < 0.21$), giving an overall positive rating for construct validity. Finally, for known groups validity, all measures indicated a difference between speakers with mild and moderate aphasia with the exception of the Local Coherence measures. Overall, Story Grammar, Reference Chains, and Predicate Argument Structure emerged as the strongest measures in the current study, as they achieved the predetermined thresholds for quality in terms of each of the psychometric parameters profiled, for both Cinderella and Everyday discourses.

Discussion and conclusion: The current study is the first to psychometrically profile measures of discourse in aphasia. It contributes to the field by identifying that Story Grammar, Reference Chains, and Predicate Argument Structure are the most psychometrically robust discourse measures profiled to date with speakers with aphasia. Until further data are available indicating the strength of other discourse measures, caution should be applied when using them.

WHAT THIS PAPER ADDS

What is already known on this subject. Very little is known about the psychometric properties of discourse measures for people with aphasia. Psychometric quality of measures is important because findings and conclusions can only be as strong as the measures that they are based upon.

What this study adds. The current study is the first study to profile the psychometric properties of acceptability, reliability, and validity of five discourse measures, using data from speakers with aphasia. The strongest measures profiled in this study were Story Grammar, Reference Chains, and Predicate Argument Structure.

Clinical implications of this study. Story Grammar, Reference Chains, and Predicate Argument Structure are currently the measures with the strongest known psychometric properties. Caution should be applied when using other discourse measures until further information is available.

MAIN TEXT

Background

Discourse is commonly defined as any unit of language above a single sentence, used for a specific purpose (Halliday 2004). Discourse is therefore the crucial unit of analysis for language used within a range of everyday interactions (Davidson *et al*. 2003), including conversation. Within the course of a naturally occurring interaction, a speaker may produce a range of monologic discourses, for example: describing the area they live in in detail; telling a story about a specific incident which happened to them; giving instructions about how to complete a credit card booking on a website; or arguing their preference for one political party over another. Each of these discourses is likely to follow a structure that can be anticipated and recognised by a listener, and which supports the listener in tracking and comprehending the content of the discourse.

The production of discourse that is structured appropriately is likely to be a complex and multifaceted process. The most comprehensive model of discourse production, reflecting this complexity, is outlined by Sherratt (2007). Sherratt describes a multilevel process (figure 1), interacting with cognitive and social pragmatic skills, where discourse production is the end point of an ‘input trigger’, shaped via cognitive and linguistic filters. Speakers first identify a discourse to communicate (input trigger); then select the shape of their discourse (frame/schema generation), then insert and integrate information from semantic and episodic memory; then assign logical relationships between information (generation, selection, and chunking of propositions); then select linguistic and lexical items (linguistic encoding); and finally produce the discourse verbally. Therefore, in line with other models and theories of discourse production (Halliday 2004; Levelt 1989; Eggins and Martin 1997), the discourse that a speaker produces drives subsequent choices around information that is included, and language that is used to do this.

---figure 1 about here--

Aphasia disrupts discourse, giving some speakers limited access to words and syntax, and affecting the information speakers can communicate (Linnik *et al.* 2016). Using the World Health Organization International Classification of Functioning (World Health Organisation 2001), a discourse impairment can be described as an impairment of body function, which impacts at the levels of activity (for example, being able to tell a story) and social participation (swapping stories within an interaction in a coffee shop). Discourse is identified as a priority for therapy by speakers with aphasia, who say that they need language for functions beyond basic needs and requirements (Worrall *et al.* 2011), and is increasingly the focus of assessment and intervention research for speakers with aphasia (Bryant *et al.* 2016). However, clinicians working with people with aphasia report not feeling confident in assessing it (Rose *et al.* 2014). This lack of confidence may be, in part, due to the number of measures available, and very little clear guidance on why one might opt for one measure over another.

Measuring discourse in aphasia

In both research and clinical practice, there is a need to quantify specific aspects of communication, and to do this, high quality measures are essential. Measurement is an emerging theme across the aphasiology literature and using high quality measures is increasingly recognised as a fundamental foundation of good research and intervention (e.g., Dietz and Boyle 2017; Wallace *et al.* 2014; Wallace *et al*. 2017). This is because research findings and clinical outcomes can only be as good as the measures that they are based on. If a measure is of high quality, the data it generates and the conclusions drawn from this data can confidently be considered robust. By contrast, if a measure is of poor quality, any data generated using the measure are likely to be questionable and resulting conclusions will be fragile. For example, if a measure has poor validity, a client's impairment may not be recognized, poorly described, or over-diagnosed. Similarly, if a measure has poor reliability, a clinician or researcher may erroneously conclude that a particular therapy is not effective; or conversely conclude that a therapy is effective when in fact the client's communication has not changed. Each of these negative outcomes impact on clinicians, researchers, and clients with aphasia themselves. Clinicians and researchers therefore have a responsibility to use measures that are of the highest possible quality, for profiling, assessment, and outcome measurement.

Psychometrics is the field concerned with the quality of measures, including properties such as acceptability, reliability, and validity. There are standard criteria for psychometric measurement (Nunnally and Bernstein 1994; Streiner and Norman 2000), providing clear and objective thresholds regarding, for example, whether a measure generates good quality data, which are complete and distributed normally (acceptability); which can be reliably scored by different scorers (inter-rater reliability), and by the same scorer over different occasions (intra-rater reliability); and which measure the targeted underlying construct (construct validity), capturing relevant and important information about it (content validity).

There is a good deal of choice when considering which discourse measures to use with people with aphasia, with recent reviews identifying more than 500 measures (Bryant et al. 2016a; Pritchard et al. 2017). This wealth of measures has been used on a range of discourses or discourse genres, such as narrative, procedural, descriptive, and personal discourses (Boyle, 2014; outlined in Bryant et al., 2016); and cover information organisation, such as measures of coherence and of story grammar; devices for making links between the information in a discourse, such as measures of cohesion; and microlinguistic measures, such as measures of sentence structure (Bryant et al., 2016a; Linnik et al. 2016; Pritchard et al. 2017). When assessing discourse, clinicians and researchers are likely to use a number of discourse measures together to describe multiple aspects and levels of discourse (Bryant et al. 2016b). Furthermore, these discourse measures are likely to be used and on a range of discourse genres, including narrative discourse, such as Cinderella discourses, which are widely the focus of discourse research (Bryant et al., 2016a), and everyday discourses, such as descriptive discourses, which might be used to describe a scene observed, or a beautiful view on holiday; personal discourses or recounts, which might be used to describe to events from the weekend to a colleague; and procedural discourse, which might be used to give instructions, for example, telling a family member about how to use a computer programme. Therefore, two key genres of discourse that clinicians and researchers seek to elicit are narrative discourse, and everyday discourse. Such everyday discourses might be elicited clinically, using prompt materials or questions. Speakers' performance on a picture description is likely to align with the everyday description tasks, such as describing a domestic scene or situation to a caller on the telephone; performance on procedural discourses is likely to align with the everyday task of giving extended instructions, such as describing how to operate a programme on a computer; and 'personal discourses' are likely to align with the everyday task of recounting a story, such as telling a colleague about something that happened over the weekend. Eliciting these using prompts is widely practised across the aphasiology research, and ensures that discourses produces are as comparable as possible across speakers. Such *everyday discourses are likely to differ from narratives in key ways: they are not culturally embedded; are likely to rely more heavily on pragmatics; and are less likely to have been rehearsed and formulaic. A listener's expectations regarding Everyday discourses are less tightly constrained than Cinderella discourses in regard to specific information in a specific order. Therefore, it is appropriate for clinicians and researchers to consider psychometrics in both key categories of discourse.* '

Very little is known about the psychometric properties of discourse measures in either narrative or everyday discourses. Pritchard *et al*. (2017) assessed 58 discourse information measures found in the aphasia literature against the psychometric criteria listed above and found that overall, the level of psychometric information reported was low. As no discourse measure targets all aspects of discourse comprehensively, content validity of the measures was evaluated for relevance against current theory (Sherratt 2007), and the majority of measures had good content validity as they were clearly related to one or more levels of the model. However, no measure in the review included data on acceptability, and only limited data was available on test-retest reliability and inter-rater reliability. Data on test-retest reliability was reported only for 8/58 measures and only 3/8 had correlations greater than 0.80. Inter-rater reliability was reported in only 4/76 studies for 12/58 measures (>.80 for 11/12 measures). Overall, there was very little to elucidate a researcher or clinician's choice of discourse measure; or to inspire confidence in the data generated by any measure. The current study aimed to address some of the gaps in what is known about the quality of discourse measures in terms of their psychometric properties.

It is worth considering *why* psychometric information on discourse measures might be so limited. There are a number of possible explanations for this, including the fact that interest in aphasic discourse has only increased relatively recently (Bryant *et al*. 2016a). However, the sparsity of information and consideration of these measures is also likely to be due to the nature of the data. In some fields, data may be straightforward to collect, generate, and analyse. For example, surveys involving self-rating questionnaires generate data that are quick to collect and to analyse, meaning that the quality of a measure can be ascertained in a relatively straightforward manner. By contrast, using a discourse measure is usually a multistep exercise (see Table 1), involving significant time commitment (Armstrong, *et al*. 2007; Boles 1998; Bryant *et al*. 2016a; Togher 2001), and using analytical skills with which clinicians and researchers may not be confident (Rose *et al*. 2014). Therefore, the complex process of generating data using discourse measures may go some way to explaining the paucity of psychometric information available.

----------------------------table 1 here--

To date, there is no information available regarding the psychometric properties of discourse measures in narrative and everyday discourses, and there is currently no clear rationale for opting for one specific approach over another. Therefore, the current study aimed to explore the psychometric properties of a range of discourse measures, selected to reflect the different levels of discourse that clinicians are likely to be focusing upon: structure, coherence, cohesion, and language (Bryant *et al*. 2016b). This was completed to ascertain their quality in narrative and everyday discourses, in terms of the following research questions:

1) What is the acceptability of each measure, in terms of % of missing data and score distribution?
2) What is the inter- and intra-rater reliability of each measure?
3) What is the content validity of each measure and what is the construct validity of each measure, in terms of correlating with related measures (convergent validity), not correlating with unrelated measures (discriminant validity) and differentiating between known groups (know-groups validity)?

Methods

Participants

Ethical approval for the study was given by the XXXXX[1]. Participants were recruited from community groups using a flyer.

Inclusion and exclusion criteria

Eligible participants were those who presented with chronic aphasia (defined as more than 6 months post stroke), and used spoken language as their primary form of communication. Participants were excluded if they had additional neuropsychological or cognitive impairments, depression, or hearing and vision that were not corrected using glasses or hearing aids.

Participants who indicated an interest were screened, using: a questionnaire for self-reported cognitive impairments and non-corrected hearing and vision; the Geriatric Depression Scale (Brink, Yesavage, Lum, Heersema, Adey and Rose 1982; Yesavage 1988) for depression, where participants were required to score between 0 - 4 out of 15 to indicate no depression; and Raven's Coloured Progressive Matrices (RCPM) (Raven, Court,

[1] To be entered after blind review has been completed.

and Raven 1995). The RCPM was used as a proxy for non-verbal cognition and it was selected as it is quick to complete, limiting participant burden; has strong psychometric properties; and does not require language production or comprehension. Participants were required to score 23 and over out of 37 to indicate non-impaired cognition, based on scores from neurologically healthy aging speakers (Basso, Capitani & Laiacona, 1987).

Measures.

Profiling measures: To profile participants' language, individuals completed a battery of language assessments, including the Western Aphasia Battery Revised- WAB-R (Kertesz 2007) to assess overall profile of language and aphasia subtype; list A of An Object and Action Naming Battery (Druks and Masterson 2000) to assess naming; and the Pyramids and Palmtrees Test (Howard and Patterson 1992), and the Kissing and Dancing Test (Bak and Hodges 2003), to assess nonverbal semantics. Aspects of these measures were also used for convergent and discriminant validity testing as described below.

Validity measures: To assess convergent construct validity, participants' scores were used from two assessments: the 'Spontaneous Speech' subsection of the WAB-R, and total score from the Kissing and Dancing Test, which we hypothesised to use some of the same skills required for spontaneous discourse production. The WAB-R spontaneous speech score assesses the utterance-level language and information participants produce when describing the picture, and the Kissing and Dancing Test assesses participants' non-verbal semantics for actions. These skills are likely to be key to discourse production. However, whilst these skills are likely to be used in discourse, neither test entirely reflects the spontaneous information and language use that can be measured using discourse measures. To assess discriminant validity, participants' scores from two assessments were used: the Auditory Verbal Comprehension subsection of the WAB-R, and the total score from the RCPM, as these measures tap underlying constructs that are related to discourse production but are nonetheless different (comprehension, cognition). In addition to hypothetical relationship, these scores from screening and profiling measures were used to test validity in order to limit participant testing burden.

Discourse samples

All participants produced a range of monologic discourses. These included the fictional story of Cinderella; and everyday discourses (*descriptive*, two picture descriptions; *procedural*, two procedural discourses; and *recount or personal* nine personal discourses, using the autobiographical incident memories from the Autobiographical Memory Interview (Kopelmann 1990) (see Appendix 1, table A1). For further detail on each of these discourses and elicitation, see Appendix 1. Discourses were elicited in a random order, completed using a random number generator, to limit order effects. Before starting the discourse samples, participants were told that they would not be interrupted or stopped. Whilst the participants were talking, the researcher offered supportive but neutral encouragement, such as smiling and nodding. A speaker was judged to be finished producing their discourse when they gave an explicit reference to being finished, for example, saying ‘and that’s everything’, or stopped talking for 10 seconds or more.

Participants completed the assessments and discourse samples across six sessions, lasting approximately one hour each. Session length was guided by participant fatigue. Participants were offered regular breaks and monitored for signs of fatigue.

Transcription and discourse analysis

Participants’ discourse samples were audio recorded. Verbatim orthographic transcriptions of the narratives were produced from these recordings, using broad phonemic transcription where appropriate. Transcription was completed via an initial first parse where an orthographic transcript was produced from the audio recording. The transcript was then checked by a trained analyst listening to the recording whilst reading the written transcript, and errors were identified, and resolved through discussion. An initial check on the accuracy of this revealed that transcription accuracy was high (99.98%), with all errors resolved through discussion. All discourse analysis was completed by hand, and each measure was used on each discourse sample.

Due to the range and breadth of discourse measures available (Bryant *et al*. 2016a; Pritchard *et al*. 2017), it is not possible to test the full range of discourse measures available; and nor is there any clear justification for why one might opt for one measure rather than another. Discourse measures were therefore selected to reflect a range of theoretical approaches, reflecting macro and microlinguistic features.

The measures used, and processes used in analysis are summarised below.

1. *Story Grammar* (based on Ulatowska *et al*. 1983). The 'base unit' for this was meaning rather than linguistic form, so Story Grammar elements can be words, phrase, sentences or longer chunks. This analysis involved identifying and tallying the story grammar elements speakers included from a total of nine: abstract (summary or introduction of the whole discourse); time (reference to the time the discourse takes place, e.g., the time of day, the season, the year); location (reference to general or specific location of the action, e.g., 'in the UK', 'in the front room'); participants (introduction of participants or objects in the discourse); a complicating action, (event causing another to occur); an event or sequence of events; evaluation: (explicit reflection on the content of the discourse, e.g., 'I left school, which was silly); result or resolution: (an event terminating the event sequence); coda (a statement signalling the discourse has finished e.g., 'so we lived happily after that'). Some elements therefore reflected using a range of linguistic base units, for example 'participants' could be introduced using single words ('Cinderella', 'the prince'), or phrases ('the lovely girl Cinderella' 'the handsome young prince'), or whole sentences ('there was a lovely girl called Cinderella').

2. *Topic Coherence*. Topic coherence was measured, using the scoring method from Mackenzie *et al*. (2007), adapted in Brady *et al*. (2003), from Mentis and Prutting (1991). This framework describes the amount a speaker divides topics into subtopics, sub-subtopics, and sub-sub-subtopics, where Topic is a clause or noun phrase that identified the question+of immediate concern and that provided a global description of the genre of a sequence of utterance; Subtopic is 'related to and germane to the main topic sequence, subordinate to the main topic in terms of being an elaboration or expansion of one aspect or dimension of the main topic'; Sub-subtopic: similar to the classification of subtopic (above), sub-subtopics are subordinate to the subtopic sequence, functioning as an 'elaboration' or 'expansion'; Sub-sub-subtopic: an additional layer expands an element of the subsubtopics. From this, a degree of topic subdivision score was calculated ((total subtopics + sub-subtopics+ sub-subsubtopics)/ total topics + subtopics+ sub-subtopics+ sub-sub-subtopics).
3. *Local Coherence* (Glosser and Deser 1990). The discourse was divided into base units of t-units (defined as a clause plus all of its dependent or relative clauses, Hunt 1970, p4). The t-units were rated on a 5-point local coherence scale, defined as the relationship between the meaning or content of a verbalization and that in the immediately preceding utterance produced either by the interviewer or by the subject. Local coherence included relationships of continuation, repetition, elaboration, subordination, or coordination with the topic in the immediately preceding verbalization (Glosser and Deser 1990, p74). Ranging from a score of five, where the topic of the preceding utterance was continued (e.g., elaboration, temporal sequencing, maintaining the same actor, subject, action, or argument as the focus); to a score of one (no relationship to the immediately preceding utterance).
4. *Reference Chains* (Marangolo *et al.* 2014). Endophoric reference chains was identified and tallied. This was completed by 1) identifying all items which featured on

a reference chain; 2) identifying which reference chain they belonged to; and 3) counting the total number of reference chains in the discourses.

5. *Predicate Argument Structure (PAS)* (Cruice, *et al.* 2014), involving identifying the main verbs in each utterance, and the arguments around each of them. Examples of 0, 1, and 2 argument structures are given in Table 2. A PAS complexity score was then calculated using the formula (number of arguments/ number of main verbs). Predicate argument structure was used to analyse the number of arguments a verb had. Only the internal arguments of a verb were counted (Cruice *et al*. 2014; Edwards *et al.* 1993; Pritchard *et al*. 2015), to accommodate omission of the subject noun phrase which is common and acceptable in the discourse of healthy speakers (e.g., within the utterances 'we went to Fairlands Park/ and ate lunch'). This meant that verb structures were calculated as having 0, 1, and 2 internal arguments (table2. Based on this, a mean predicate argument structure score was calculated (Cruice *et al*. 2014; Webster *et al*. 2007), using the calculation (total number of arguments produced/ total number of predicates produced). Non-arguments or adjuncts were removed from analysis – these are non-core semantic arguments relating to aspects such as place, time, measure, accompaniment (Black and Chiat 2003). For example, the phrase 'at the weekend' in the utterance [He] went [on at date] (at the weekend), would be classified as a non-argument.

------------------------table 2 about here---

Psychometric assessment of measures

As indicated above, discourse measures were evaluated for acceptability, reliability and validity. Table 3 details the criteria against which each property was assessed. For acceptability, we evaluated the completeness of the data in terms of missing values and score distribution. For reliability, we evaluated inter-rater reliability by calculating correlations between scores of different raters on the same discourse measure (the first author of the current study and a trained Speech and Language Therapist, using 20% of the data); and intra-rater reliability by calculating correlations of scores by the same rater on the same discourse measure, when viewed on different occasions (the first author of the current study, using 20% of the data, scored 12 weeks apart). To minimize participant burden, test-retest reliability was not completed for the current study. Content validity was evaluated by qualitatively evaluating each measure against the Sherratt (2007) theoretical model of discourse production, to determine whether it clearly reflected one or more than one stage of production. Three aspects of construct validity were assessed. For convergent validity, we hypothesized that measures of discourse will have low - moderate correlations with the Kissing and Dancing Test and the Spontaneous Speech Subsection of the WAB-R. We did not anticipate high correlations with these reference measures as their scores largely related to utterance level language and information, rather than the discourse level under focus in the current study. For discriminant validity, we hypothesized that measures of discourse will have negligible or no correlations with less related constructs, the RCPM and the Comprehension Subsection of the WAB-R. Lastly, for known-groups validity we hypothesized that people with mild aphasia will have better discourse scores than people with moderate or severe aphasia.

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

Data analysis

Participants' scores were analysed as two groups: Cinderella discourses, and Everyday discourses (collapsed data from personal, procedural, and picture description discourses). The purpose of this grouping of the data was to differentiate Cinderella discourses, which are the most widely used narrative discourse elicitation method (Bryant et al. 2016a), particularly in analysis of Predicate Argument Structure (e.g., Berndt *et al*. 1997; Saffran *et al.* 1989; Webster *et al.* 2007), and everyday discourses.

Scores for all measures were collated into a SPSS spreadsheet (IBM 2016). Descriptive statistics were used to describe participants and generate distributions of measures' scores. Correlations were explored between discourse measures to see if they tapped different underlying aspects of discourse (r < .60) or whether there was substantial overlap between them suggesting redundancy. To evaluate acceptability, the completeness of data (missing data <10 %) and the normality of data distribution (skewness values between -1 and +1) were assessed. Reliability was assessed by calculating intra-class correlation coefficients (ICCs), using a two-way random effects model. ICCs had to be > .80 for good inter- and intra-rater reliability. Content validity was evaluated qualitatively by checking the relevance of each discourse measure against the Sherratt theoretical model. For convergent validity, Pearson's correlations *r* had to be moderate, i.e. r > .30 between measures of discourse and related measures; and for discriminant validity r had to be < .30. To assess known-groups validity, participants were split into aphasia severity groups based on the WAB-R Aphasia Quotient thresholds. We had two groups, AQ = 48 - 75 moderate aphasia, and AQ > 75 mild aphasia. We hypothesized that the mild aphasia group will have higher discourse scores for all measures than the moderate aphasia group. Given the small sample size, differences in scores were evaluated with Cohen's *d* (Cohen, 1988) effect sizes, using the formula *d* = (m1-m2)/ SD Pooled, where SD pooled was calculated using the formula (SD1+SD2)/2. *d* was interpreted using Cohen's thresholds of *d* = .2 small, *d* = .5 medium, and *d* = .8 large effect size.

Results

Data were collected from 17 participants, aged 25-73 (M= 53.1, SD= 13.1). There were five females and 12 males. All participants were right handed, and all reported single left hemisphere strokes. Participants' years of education ranged from 10- 23 (M= 14.41, SD= 3.72), and months post stroke ranged from 24-109 (M= 70.12, SD= 41.92). Table 4 details participants' demographic characteristics.

------------------------------------table 4 about here--

Screening, profiling and validity measures

All participants reported normal hearing and vision, and no participant obtained scores indicating depression on the *GDS*. Scores on the *GDS* ranged 0 – 4 with a mean (SD) of 1.06(1.3). No participant reported any neuropsychological impairment additional to aphasia.

Each participant scored within normal limits on *RCPM*. Scores on the *RCPM* ranged 24- 32 with a mean (SD)= 25.94(2.29).

Participants presented with a range of mild to moderate aphasia profiles, as defined by the WAB-R, with a mean (SD) AQ = 76.93 (15.51), (range = 49.6 - 93.8). See table 5 for a further breakdown of participants' scores on screening, profiling and validity measures.

--table 5 about here----------------------------------

Discourse measures

Descriptive statistics were inspected for each discourse measure (table 6). Correlation analysis between the different discourse measures (table 7) indicated that there was no redundancy between them.

--table 6 about here------------------------------

---table 7 about here-----------------------------

RQ 1) What is the acceptability of each measure, in terms of % of missing data and score distribution?

All measures yielded complete data, suggesting that each could be used on a broad range of discourses produced by participants with aphasia. No measure yielded skewed data (table 6).

RQ 2): What is the inter- and intra-rater reliability of each measure?

All measures demonstrated high levels of intra-rater reliability (ICC = .92 - .97) and inter-rater reliability (ICC = .9 - .95) (table 8).

----------------------------table 8 about here--

RQ 3): What is the validity of each measure, in terms of relating to a theoretical model of discourse production (content validity), correlating with related measures (convergent validity), not correlating with unrelated measures (discriminant validity) and differentiating between known groups (know-groups validity)?

For content validity, we found that each measure related to one or more levels of the theoretical model of discourse production by Sherratt (2007) (figure 1). The measure of Story Grammar, is a measure which quantifies the specific information components in a discourse. It therefore reflects the overall discourse frame, relates to the stage of 'frame/ schema generation' on the model of Sherratt (2007). The measures of Topic Coherence, Local Coherence, and Reference Chains relate to the stage of 'generating and chunking propositions' because they are measures of information organisation and links between utterances. The measure of Predicate Argument Structure relates to the stage of 'linguistic encoding', as a clause or utterance- level measure.

For convergent validity (table 9), focusing on correlations with the Spontaneous Speech Subsection of the WAB-R, 2/ 10 measures had low correlations (Cinderella Topic Coherence and Cinderella Local Coherence); and 8/10 had correlations of r > .30 as expected, with one of these measures (Everyday Topic Coherence) having a high correlation (r = .76). In terms of correlations with the Kissing and Dancing Test, most were moderate as expected (7/10 measures r > .30), with the Everyday Topic Coherence having the highest correlation (r = .58). Cinderella Local Coherence, Cinderella Predicate Argument Structure, and Everyday Story Grammar had correlations of r = .26, r = .27 and r = .26 respectively. This means that overall, 15/20 (75%) of the measures met the threshold for convergent validity.

For discriminant validity (table 9), as expected, no measure had a correlation of r > .30 with either scores from the RCPM (0/10), or with the Auditory Verbal Comprehension Score from the WAB-R (0/10). This meant that overall, 20/20 (100%) of measures met the threshold for discriminant validity.

---table 9 here---------------------------------------

For known groups validity, descriptive statistics indicated a difference between the groups of mild and moderate aphasia, with the group with mild aphasia scoring more highly than the moderate group across all measures (table 10). There was a very small effect size (*d* = .04 - .17) for 2/10 measures (Cinderella and Everyday Local Coherence); a small – medium effect size (*d* = .28 - .64) for 7/10 measures (Cinderella Story Grammar, Cinderella Topic Coherence, Cinderella Reference Chains, Cinderella PAS, Everyday Story Grammar, Everyday Reference Chains, Everyday PAS); and a large effect size (*d* = .9) for 1/10 measures (Everyday Topic Coherence).

------------------------------table 10 about here---

In summary, all measures scored highly for the criteria of acceptability, reliability (intra- and inter- rater), content validity, and discriminant construct validity. For convergent construct validity, the Topic Coherence and Local Coherence measures for the Cinderella discourses did not meet the criterion set for one of the two formal assessment measures (the Speech Subsection of the WAB-R). For the construct validity- known groups, the Local Coherence measure did not demonstrated differences between those with mild and those with moderate aphasia for both Cinderella and Everyday discourses. Therefore, the measures that passed all set criteria for both Cinderella and Everyday discourses were Story Grammar, and Reference Chains, and mean Predicate Argument Structure complexity; and the measures which appear weaker are the measures of Topic Coherence and of Local Coherence.

Discussion

The current study aimed to further the field of discourse analysis in speakers with aphasia, using data from a group of people with mild to moderate aphasia (n=17), to investigate the psychometric properties, in terms of the acceptability, validity and reliability of a set of key discourse measures. Using high quality measures is of paramount importance to good research and outcome measurement (Wallace *et al*. 2014), including assessment and intervention with people with aphasia, as the strength of findings and conclusions of any work relies on the quality of measures. The measures in the current study were selected to reflect a range of features of discourse that clinicians are likely to measure, and to include four macrolinguistic measures (Story Grammar, Topic Coherence, Local Coherence, and

Reference Chains) and one microlinguistic measure (PAS), and which collectively target story structure, coherence, cohesion, and sentence structure. This is the largest study of its kind completed to date, and strengthens the field by furthering what is known about the quality of these measures, informing clinical work and research work in the field of discourse analysis. Overall, the strongest measures in the current study were Story Grammar, Reference Chains, and Predicate Argument Structure, as these were the only measures which demonstrated quality across each of the domains of acceptability, reliability, and validity, with other measures not demonstrating quality thresholds for all of the domains.

We are not aware of any previous study which has explicitly reported data on the acceptability of discourse measures, and the review by Pritchard *et al*. (2017) identified this as a gap in the evidence base relating to discourse measurement. Whilst acceptability data itself may appear straightforward, it is seldom reported. Explicitly reporting such data is important, as such data reflect the suitability of a measure in a given context. In the current study, there were no missing data, meaning each of the measures was suitable for use on a range of discourses produced by speakers with mild- to moderate aphasia, as defined by the WAB-R (Kertesz 2007). Furthermore, none of the data were skewed (skewness range= -.33 - .47), suggesting an approximately normal distribution. This is of interest, as normal distribution is a core assumption regarding naturally occurring variables. In terms of statistical comparisons, an assumption of normal distribution of data underpins a large number of statistical tests routinely used to identify differences between groups. Parametric tests are used to identify differences between groups of speakers with aphasia; and are used in intervention research, to describe changes which occur after therapy. The data from the current study suggests that the use of such tests with discourse data is likely to be appropriate, even with small to moderate sample sizes of 15-20 participants as in the current study.

The current study is the first study we are aware of to look at both inter-rater and intra- rater reliability of five discourse measures, using reliability statistics. Previous research has not used robust means to assess reliability such as kappa and the Intraclass Correlation Coefficient (ICC) (Hallgren 2012), with the majority of reliability data for discourse measures relying on percentage agreement (Pritchard *et al*. 2017). The current study suggests that each of the five measures have very high levels of reliability (ICC≥ .9), both for different raters using a measure, and the same rater at different time points. Such a finding suggests discourse measures can be used in assessment and therapy settings, where it is a necessity that the same measure can be used consistently by research and clinical professionals. Such a finding regarding reliability is perhaps surprising, given that clinicians report not feeling confident in this area of assessment (Rose *et al.* 2014). There are two possible explanations for this finding. The first is that clinicians are skilled in discourse analysis and simply lack confidence; and the second is that the training given to clinicians conducting the analysis in the current study increased the reliability. Although the data from the current study cannot speak directly to these two possibilities, the majority of studies of discourse measurement in aphasia report high levels of reliability in trained analysts. It is possible and indeed likely that training increases confidence and skills in discourse analysis, and training is likely to be key in discourse measurement across clinical and research settings.

In terms of content validity, similar to the findings of Pritchard *et al*. (2017), each of the measures in the current study could be related to the theoretical model of discourse production (Sherratt 2007), reflecting a similar underlying concept. This is important, because it supports clinicians and researchers to identify the level at which discourse is impaired. Although it is likely that the majority of discourse measures can be related to such a model (Pritchard *et al*. 2017), a link between the discourse measure and underlying stages of discourse production is not often made explicit in research articles, and consequently it is not always clear why a particular measure has been selected from the abundance of discourse measures which are available (Bryant *et al*. 2016a). The consistent and explicit linking of discourse measures to underlying models is likely to clarify clinicians' and researchers' rationales for selecting specific measures, and to inform this decision-making process.

When we looked at convergent validity, the majority of discourse measures met our expectations, with 75% having a moderate correlation with the measures of spontaneous speech and non-verbal semantics, in line with positive thresholds for hypothesis- based construct validity in published literature (Terweea, Bota, de Boera, van der Wind, Knola, Dekkera, Boater& de Veta, H., 2007). This suggests that discourse measures reflect similar skills to those examined within related assessments. The majority of these measures had low – medium correlations (r = .25- .5), suggesting that whilst there is some relationship between the constructs reflected within published language assessment scores and scores generated by discourse measures, there are other factors at play. This might be predicted when considering theoretical models of discourse production: the majority of published assessments assess discrete aspects of communication and language, such as non-verbal semantics, sentence production, or ability to communicate discrete units of information. It is highly likely when considering models such as that described by Sherratt (2007), that discourse is a multifaceted task, drawing simultaneously on multiple linguistic and cognitive processes. Therefore, any published language assessment measure is likely to reflect only one or a small combination of these facets. The low to medium correlations found in the current study between the scores on the published assessments and discourse measures strengthen the case for measurement of discourse in addition to completing published assessments. This low to moderate relationship between the published assessments and the discourse measures suggests that if a clinician or researcher uses only one or the other, they may not fully profile a speaker's skills.

The only measures which had very low correlations for our assessment of convergent construct validity were the discourse measures of Topic Coherence and Local Coherence (for the Cinderella discourse), with the Spontaneous Speech subsection of the WAB-R. It is likely that this is because in these contexts, the skills a speaker is scored on are very different. In the WAB-R Spontaneous Speech subsection, the score reflects the assessor's subjective judgment of whether a speaker communicates in fluent, syntactically appropriate sentences, and communicates specific information, including items and activities. By contrast, the scores on the Topic Coherence and Local Coherence measures in a Cinderella story reflect whether or not a speaker produces a sequence of utterances which clearly relate to each other (Local Coherence), and which are hierarchically organised (Topic Coherence). Therefore, whilst both are measures of discourse, they are likely to reflect different underlying constructs and skills. These measures did correlate for Everyday discourses. This is likely to be because this group of discourses included picture description (like the WAB-R) which do not require hierarchical organization of information (Topic Coherence) and maintaining coherence from one utterance to the next (Local Coherence). Those discourses in this category which were not picture description (procedural discourses and personal discourses) also required less hierarchical organization and coherence than the Cinderella discourse, as they were not culturally embedded narrative discourses. Therefore, it is likely that whilst the Everyday discourse tapped similar skills to the WAB-R Spontaneous Speech subsection, the Cinderella discourses did not.

For discriminant validity, in line with our expectations, 100% of the discourse measures had low correlations ($r < .3$) with the measures of the auditory verbal comprehension and cognition as measured by the RCPM, giving a positive rating for construct validity based on the thresholds decribed by Teerwe et al. (2007). This by no means suggests that cognition and comprehension are not important in discourse. Rather, it suggests a small overlap with these constructs, as a number of processes are also at play to form a good discourse (Sherratt 2007).

Finally, when considering known groups validity, the majority of measures indicated a difference between speakers with mild and moderate aphasia. This finding is in line with research which indicates that discourse is a key indicator of aphasic difficulty, and that this is different depending on the severity of a speaker's aphasia (Fromm *et al.* 2017). Only the measures of Local Coherence indicated no meaningful difference between the groups. This may be because the particular measure of Local Coherence used in this study is insufficiently sensitive to reflect the difference in the local coherence of discourse with speakers with mild and moderate aphasia. By contrast, in the Everyday discourses, the measures of Story Grammar and Topic Coherence indicated a moderate and large effect size, respectively. These suggest that these factors (the amount of relevant information a speaker includes in discourse, and how they organise that information) are key factors discriminating between speakers with mild and moderate aphasia.

Limitations and Future Research

The current study assessed the psychometric properties in terms of acceptability, reliability and validity of five discourse measures only, and with a limited participant sample. There a large number of discourse measures reported within the literature (Bryant et al, 2016; Pritchard et al., 2017), reflecting a broad range of theoretical bases. Future research should aim to further profile a wider range of discourse measures, including detailed exploration and articulation of theoretical underpinnings of the measures, and using data from larger groups of speakers with aphasia. Furthermore, the current study identified that the two measures of information in discourse, relating to the level of 'generating and chunking propositions' (Sherratt 2007, p377) were not of high quality across all psychometric domains. This suggests that further work on discourse measures and their quality is warranted, to identify a psychometrically higher quality measure relating to information at this level of the model ('selection and topicalisation of information', and 'generation and chunking of propositions', Sherratt 2007, p377). Finally, grouping discourses together for investigation as 'everyday discourses' may mean that the variability of individual categories is not explored. Future psychometric profiling work should expand this by focusing on profiling different everyday discourse genres.

Clinical Implications

Story Grammar, Reference Chains, Predicate Argument Structure emerged as the strongest measures profiled in the current study. Caution should be applied when using other discourse measures until further information is available.

References

ARMSTRONG, L., BRADY, M., MACKENZIE, C., and NORRIE, J., 2007, Transcriptionless analysis of aphasic discourse: A clinician's dream or a possibility? *Aphasiology,* **21(3/4)**, 355- 374

ARMSTRONG, E., GODECKE, E. and KOK, B., 2011, Monologues and dialogues in aphasia: some initial comparisons. Aphasiology, **25**, 1347–1371

BAK, T., and HODGES, J., 2003, Kissing and dancing: A test to distinguish the lexical and conceptual contributions to noun/verb and action/object dissociation. Preliminary results in patients with frontotemporal dementia. *Journal of Neurolinguistics,* **16**, 169-181.

BERNDT, R., MITCHUM, C., HAENDIGES, A., and SANDSON, J., 1997, Verb retrieval in aphasia. 1. Characterizing single word impairments. *Brain and language*, ***56*****(1)**, 68-106.

BOLES, L., 1998, Conversational discourse analysis as a method for evaluating progress in aphasia: A case report. *Journal of Communication Disorders,* **31**, 261-274.

BRINK, T., YESAVAGE, J., LUM, O., HEERSEMA, P., ADEY, M., and ROSE, T., 1982, Screening tests for geriatric depression. *Clinical Gerontologist,* **1**, 37-43.

BRYANT, L., FERGUSON, A., and SPENCER, E., 2016a, Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical linguistics and phonetics*, **30(7)*,*** 489-518.

BRYANT, L., SPENCER, E., and FERGUSON, A., 2016b, Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology*, **31(10)**, 1105-1126.

COCKS, N., HIRD, K., and KIRSNER, K., 2007, The relationship between right hemisphere damage and gesture in spontaneous discourse. *Aphasiology,* **21(3-4)**, 299-319.

COHEN, J., 1988, *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

CRUICE, M., PRITCHARD, M., and DIPPER, L., 2014, Verb use in aphasic and non-aphasic personal discourse: What is normal? *Journal of Neurolinguistics*, **28**, 31-47.

DAVIDSON, B., WORRALL, L., and HICKSON, L., 2003, Identifying the communication activities of older people with aphasia: Evidence from naturalistic observation. *Aphasiology*, **17(3)**, 243-264.

DIETZ, A., and Boyle, M., 2017, Discourse measurement in aphasia research: Have we reached the tipping point? *Aphasiology*, early online, doi:10.1080/02687038.2017.1398803

DRUKS, J., and MASTERSON, J., 2000, *Object and Action Naming Battery*. Hove: Psychology Press.

EGGINS, S., and MARTIN, J., 1997, Genres and registers of discourse. In van Dijk, T. (Ed.), *Discourse as structure and process: Discourse studies: A multidisciplinary introduction*. London: Edward Arnold.

FROMM, D., FORBES, M., HOLLAND, A., DALTON, S. G., RICHARDSON, J., and MACWHINNEY, B., 2017, Discourse Characteristics in Aphasia Beyond the Western Aphasia Battery Cutoff. *American Journal of Speech-Language Pathology*, **26**, 762-768.

GLOSSER, G., and DESER, T., 1990, Patterns of discourse production amoung neurological patients with fluent language disorders. *Brain and Language,* **40**, 67-88.

HALLGREN, 2012, Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology,* **8(1)**, 23–34.

HALLIDAY, M., 2004, An introduction to functional grammar. London: Hodder Education.

HOWARD, D., and PATTERSON, K., 1992, *Pyramids and Palmtrees*. Oxford: Pearson Assessment.

IBM, 2016, IBM SPSS Statistics for Mac, Version 23.0. Armonk, NY: IBM Corp.

KERTESZ, A., 2007, *The Western Aphasia Battery WAB-R*. London: Grune and Statton.

RAVEN J., COURT J., and RAVEN J., 1995, Coloured Progressive Matrices. Oxford: Oxford Psychologists Press

KOPELMAN, M., WILSON, B., and BADDELEY, A., 1990, *Autobiographical Memory Interview (AMI)*. London: Pearson Assessment.

LEVELT, W. (1989). Speaking. From intention to articulation. *MA: The MIT Press, Cambridge.*

LINNIK, A., BASTIAANSE, R., and HÖHLE, B., 2016, Discourse production in aphasia: a current review of theoretical and methodological challenges. *Aphasiology*, **30(7)**, 765-800.

MACKENZIE, C., BRADY, M., NORRIE, J., and POEDJIANTO, N., 2007, Picture description in neurologically normal adults: Concepts and topic coherence. *Aphasiology,* **21(3-4**), 340-354.

MARANGOLO, P., FIORI, V., CAMPANA, S., CALPAGNANO, M. A., RAZZANO, C., CALTAGIRONE, C., and MARINI, A., 2014, Something to talk about: enhancement of linguistic cohesion through tdCS in chronic non fluent aphasia. *Neuropsychologia*, **53**, 246-256.

NICHOLAS, L., and BROOKSHIRE, R., 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.

NUNNALLY, J. and BERNSTEIN, I., 1994, Psychometric Theory. New York: McGraw Hill

PRITCHARD, M., COCKS, N., DIPPER, L., and MORGAN, G., 2015, Aphasia, procedural narrative, and iconic gesture. *Aphasiology*, **29**, 826–844.

PRITCHARD, M., HILARI, K., COCKS, N., and DIPPER, L., 2017, Reviewing the quality of discourse information measures in aphasia. *International Journal of Language and Communication Disorders,* **52(6)**, 689- 732.

ROSE, M., FERGUSON, A., POWER, E., TOGHER, L., and WORRALL, L., 2014, Aphasia rehabilitation in Australia: Current practices, challenges and future directions. *International journal of speech-language pathology*, **16(2)**, 169-180.

SAFFRAN, E., BERNDT, R., and SCHWARTZ, M., 1989, The quantitative analysis of agrammatic production: Procedure and data. *Brain and language*, **37(3)**, 440-479.

SHERRATT, S., 2007, Multi- level discourse analysis: A feasible approach. *Aphasiology*, ***21*****(3-4)**, 375-393

STREINER, D. and NORNAM, R., 2000, *Health Measurement Scales: A Practical Guide to Their Development and Use.* Oxford: Oxford University Press

TERWEE, C., BOTA, S., DE BOERA, M., VAN DER WIND, D., KNOLA, D., DEKKERA, J., BOATER, L., & DE VETA, H., 2007, Quality criteria were proposed for measurement properties of health status questionnaires. Jo*urnal of Clinical Epidemiology,* **60**, 34- 42.

TOGHER, L., 2001, Discourse sampling in the 21st Century. *Journal of Communication Disorders,* **34**, 228- 244.

ULATOWSKA, H., FREEDMAM STERN, R., WEISS DOYEL, A., MACALUSO-HAYNES, S., and NORTH, A., 1983, Production of narrative discourse in aphasia. *Brain and Language,* **19**, 317-334.

VILLARDITA, C., 1985, Raven's Coloured Progressive Matrices and intellectual impairment in patients with focal brain damage. *Cortex,* **21**, 627-634.

WALLACE, S., WORRALL, L., ROSE, T., and LE DORZE, G., 2014, Measuring outcomes in aphasia research: A review of current practice and an agenda for standardisation. *Aphasiology*, **28(11)**, 1364-1384.

WEBSTER, J., FRANKLIN, S., and HOWARD, D., 2007, An analysis of thematic and phrasal structure in people with aphasia: What more can we learn from the story of Cinderella? *Journal of Neurolinguistics,* **20(5)**, 363-394.

WORLD HEALTH ORGANIZATION, 2001, *International Classification of Functioning, Disability and Health: ICF*. World Health Organization.

WALLACE, S., WORRALL, L., ROSE, T., and LE DORZE, G., 2014, Measuring outcomes in aphasia research: A review of current practice and an agenda for standardisation. *Aphasiology*, ***28*****(11),** 1364-1384.

WALLACE, S., WORRALL, L., ROSE, T., and LE DORZE, G., 2017, Discourse measurement in aphasia research: have we reached the tipping point? A core outcome set … or greater standardisation of discourse measures? *Aphasiology*, early online, doi: 10.1080/02687038.2017.1398811

WORRALL, L., SHERRATT, S., ROGERS, P., HOWE, T., HERSH, D., FERGUSON, A., and DAVIDSON, B., 2011, What people with aphasia want: Their goals according to the ICF: *Aphasiology*, **25(3)**, 309-322.

YESAVAGE, J., 1988, Geriatric Depression Scale. *Psychopharmacology Bulletin,* **24**, 709-710.

Table 1:
Steps in using a discourse measure

	Step	Description
1	Capture method	Audio/ video method of capture. The majority of discourse measures cannot be applied live.
2	Elicitation	Eliciting discourse from the client, using stimulus such as a picture, picture sequence, wordless picture book, video, or story/ event from the participants' memory (methods outlined by Bryant *et al*. 2016a).
3	Data transfer	If necessary, the recording must be transferred from a device (e.g., video camera), to be transcribed and stored.
4	Transcription	Offline orthographic transcription from listening to recorded discourse.
5	Transcript preparation(s)	For example, parsing the discourse into base units such as clauses to complete further analysis (e.g., C-Unit analysis, Armstrong, Godecke and Kok 2011), or removing specific features of the discourse, such as repetitions and fillers (QPA, Saffran, Berndt and Schwartz 1989). If multiple measures are used, and each measure uses different base units, this step is repeated for each measure used.
6	Analyses	Applying analysis to each base unit. For example, identifying the Correct Information Units (CIUs) in a discourse sample requires analysing how many words are intelligible in context, relevant, and informative (Nicholas and Brookshire 1993). If multiple discourse measures are used, this step is repeated for each measure.
7	Collating analyses	Scores from analysis may need to be summarised to generate a single numerical figure, or percentage, for example, # or % CIUs. If multiple discourse measures are used, this step is repeated for each measure.

Table 2

Examples of 0, 1, and 2 predicate argument structures as analysed in the current study

Number of internal arguments	Examples
0 argument	*[I] left* *[he] smiled* *[Jacob] was eating*
1 argument	*[I] enjoyed [that snack]* *[the kid] went [to the pool]* *[he] wanted [that one]*
2 arguments	*[I] threw [the page] [away]* *[he] put [it] [over there]* *[he] gave [the boy] [the balloon]*

Table 3:
Definitions and criteria for psychometric properties (based on Nunnally and Bernstein 1994; and Streiner and Norman 2000)

Psychometric property		Definition	Quality criteria
Acceptability		Overall quality of the data, assessed by completeness of the data and score distribution	Missing data <10 % Skewness between -1 and +1
Reliability	Intra-rater reliability	Stability of a measuring instrument on repeated administrations by a single rater.	ICC > .80
	Inter-rater reliability	Two or more raters agreeing in their judgement using a measure	ICC > .80
Validity	Content validity	Evidence that measures are consistent with a conceptual model.	Evaluated qualitatively by checking relevance of measure to a theoretical model of discourse production (Sherratt, 2007)
	Construct validity – discriminant	Low correlations with measures of unrelated constructs.	r < .30
	Construct validity – convergent	Moderate correlations with measures of similar constructs (Scores from the Spontaneous Speech of the Western Aphasia Battery, Kertesz, 2007; and the Kissing and Dancing Test, Bak & Hodges, 2003)	r > .30
	Construct validity – known groups	Known groups differences/ hypothesis testing	Differences (medium effect sizes) in discourse scores between different groups: speakers with mild vs speakers with moderate aphasia.

Table 4

Demographic characteristics of participants with aphasia (n=17)

#	Age	Gender	First/ Main language	Other languages	Pre-stroke occupation	Years of education	Months post stroke
1	45	F	English	Shona	Restaurant owner	13	180
2	50	M	English	-	Journalist	12	40
3	50	M	English	Arabic, French	Computer Repairs	14	37
4	55	M	English	-	Bus Driver	14	48
5	60	M	English	-	Office manager	12	61
6	73	F	English	-	Mental Health Nurse	11	53
7	57	F	English	Gallic	Artist	20	109
8	54	M	English	-	Solicitor	22	67
9	54	M	English	-	Accountant	13	40
10	73	M	English	-	Post office worker	10	120
11	71	M	English	-	Chartered Accountant	12	193
12	36	M	English	Russian, Polish	Computer Scientist	23	54
13	25	F	English	Spanish	Student	15	36
14	32	M	English	-	IT consultant	14	24
15	56	F	English	-	Charity shop worker	15	52
16	49	M	English	-	Care support worker	13	47
17	62	M	English	-	Florist	12	31
Mean (SD)	53.06 (13.05)	-	-	-	-	14.4 (3.63)	70.12 (49.05)

Table 5

Assessment scores of participants (n=17)

	Western Aphasia Battery- Revised								Object and Action Naming Battery (List A)		Semantics		Geriatric Depression Scale (Max =15)	Raven's Coloured Progressive Matrices (Max = 37)
Participant Number	Info Content (Max=10)	Fluency (Max= 10)	Spontaneous speech (Max= 20)	Auditory verbal comp (Max= 10)	Repetition (Max= 10)	Naming (Max= 10)	AQ	Aphasia Subtype	Objects (Max= 81)	Actions (Max= 50)	PPT (Max= 52)	KDT (Max=52)		
1	9	6	14	8.9	7.8	8.1	77.6	Anomia	69	34	47	49	0	24
2	9	9	18	9.4	9.4	9.4	92.4	Anomia	74	41	50	48	1	24
3	9	4	13	9.4	8.6	8.2	78.4	TC Motor	78	31	49	48	4	32
4	10	9	19	9.2	9.4	8.3	91.8	Anomia	72	43	50	51	3	26
5	5	6	11	5	3.2	5.6	49.6	Wernicke's	47	18	48	43	3	25
6	7	6	13	5.85	4.2	6.7	59.5	Wernicke's	71	35	47	47	2	25
7	10	9	19	9.8	9.4	8.7	93.8	Anomia	72	40	50	51	0	26
8	10	9	19	9.75	9.6	8.2	93.1	Anomia	67	41	52	52	0	26
9	8	9	17	9.3	9.2	9.7	90.4	Anomia	74	43	52	51	0	29
10	7	4	11	6.4	5.6	5	58	Broca's	18	6	33	32	0	26
11	9	9	18	8.5	8.6	8.3	86.8	Anomic	75	38	42	52	1	25
12	9	9	18	9.85	9.2	9.5	93.1	Anomic	79	45	50	48	0	24
13	9	9	18	8.15	6.8	8	83.5	Conduction	77	44	51	49	2	24
14	4	8	12	7	7.4	8	68.8	Anomic	48	35	48	48	0	30
15	3	9	12	8.6	7.1	7.3	70	Conduction	52	42	28	45	0	24
16	8	9	17	8.2	6	4	70.4	Conduction	31	21	49	49	1	24
17	4	8	12	9	4.3	5	50.6	Conduction	34	35	47	49	1	27
Mean (SD)	7.65 (2.22)	7.76 (1.77)	15.35 (3.05)	8.37 (1.42)	7.4 (2.02)	7.53 (1.65)	76.93 (15.05)	-	61.06 (18.42)	34.82 (10.31)	46.65 (6.38)	47.77 (4.56)	1.06 (1.3)	25.94 (2.29)

Table 6

Descriptive statistics of discourse measures (n = 17)

Discourse sample	Measure	Min	Max	Mean (SD)	Median (IQR)	Skewness
Cinderella	Story Grammar	2	5	3.47 (1.23)	3.00 (3)	.38
	Topic Coherence	0	66.67	43.39 (19.85)	50.00 (25.56)	-.33
	Local Coherence	4.34	5	4.71 (.21)	4.75 (.27)	-.19
	Reference Chains	0	7	3.11 (2.17)	3.00 (4)	.05
	Predicate Argument Structure	.33	1.09	.82 (.17)	.86 (.19)	-.24
Everyday	Story Grammar	2	5.15	2.86 (.78)	2.69 (.9)	.21
	Topic Coherence	37.22	73.81	57.71 (12.45)	58.21 (21.92)	-.04
	Local Coherence	4.26	4.99	4.62 (.22)	4.66 (.36)	-.18
	Reference Chains	.58	3.31	1.75 (.93)	1.31 (1.72)	.47
	Predicate Argument Structure	.65	1.01	.86 (.11)	.87 (.17)	-.09

Table 7

Correlations (r) among discourse measures, averaged across discourses

	Story Grammar	Topic coherence	Local coherence	Reference Chains	PAS
Story Grammar		.40	.26	.59	.25
Topic coherence			.52	.57	.23
Local coherence				.35	.14
Reference Chains					.42

Table 8

Intra- and inter-rater reliability of discourse measures

Measure	Intra- rater			Inter- rater		
	ICC	SEM	95% CIs	ICC	SEM	95% CIs
Story Grammar	.96	0.20	[.94-.98]	.94	0.25	[.90-.97]
Topic Coherence	.97	5.05	[.95-.98]	.95	6.50	[.91-.97]
Local Coherence	.92	0.28	[.86-.95]	.90	0.31	[.83-.94]
Reference Chains	.95	0.22	[.91-.97]	.93	0.26	[.88-.96]
Predicate Argument Structure	.97	0.05	[.95-.98]	.94	0.08	[.91-.97]

Table 9

Convergent and discriminant validity of discourse measures

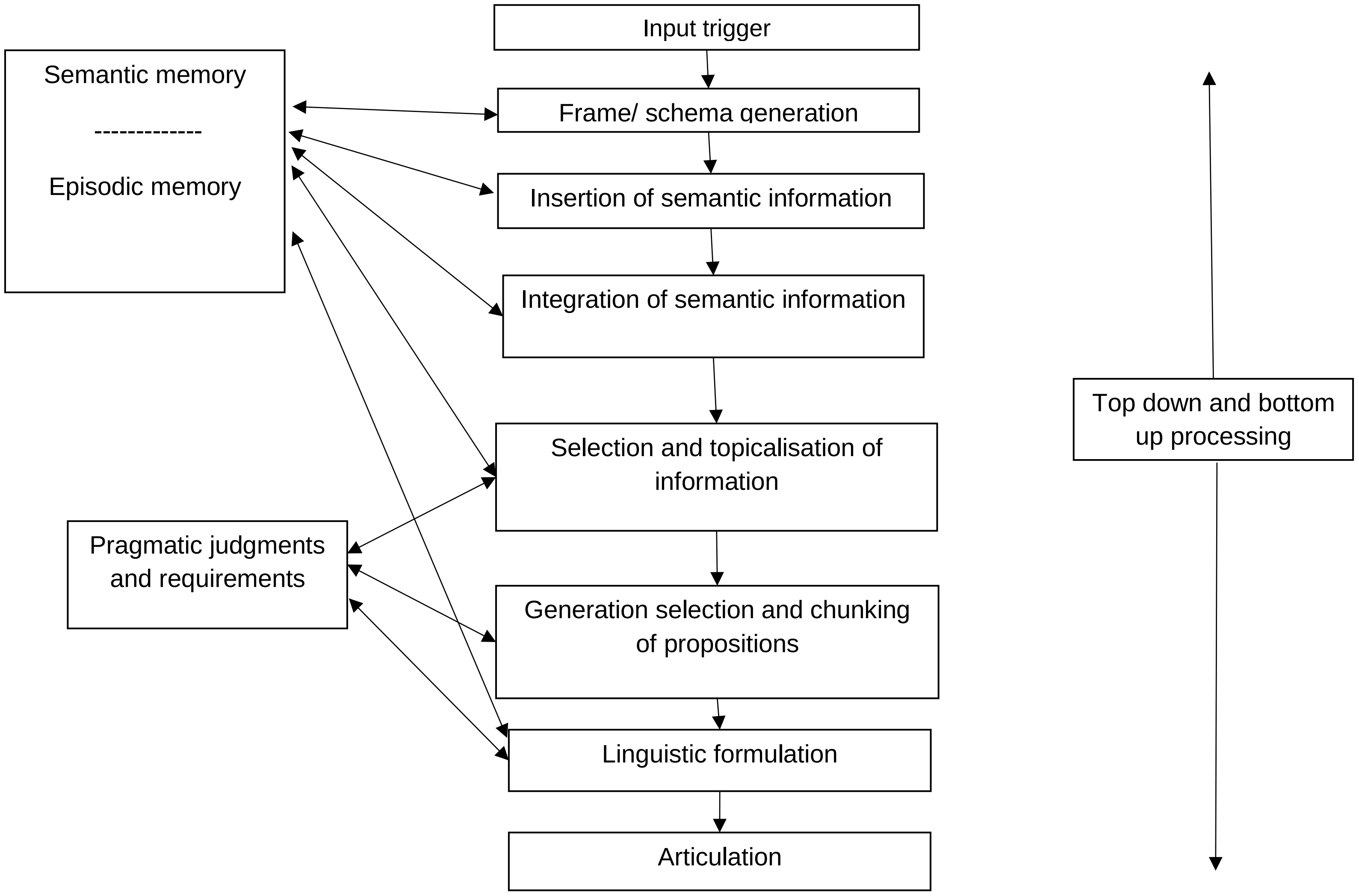
Other measures	Convergent validity									
	Cinderella					Everyday				
	Story Grammar	Topic coherence	Local Coherence	Reference chains	PAS	Story Grammar	Topic coherence	Local Coherence	Reference chains	PAS
Spontaneous Speech	.30	.23	.09	.34	.30	.30	.76*	.31	.41	.45
Kissing and Dancing Test	.45	.49*	.26	.30	.27	.25	.58*	.39	.31	.30
	Discriminant validity									
	Cinderella					Everyday				
	Story Grammar	Topic coherence	Local Coherence	Reference chains	PAS	Story Grammar	Topic coherence	Local Coherence	Reference chains	PAS
Raven's Coloured Progressive Matrices	-.05	-.13	.11	.08	.16	.18	-.16	-.02	-.10	.03
Auditory Verbal Comprehension Score	.07	-.20	-.03	.10	-.10	.03	-.08	.10	.18	.21

*Significant at the 0.05 level

Table 10

Known groups validity of discourse measures

		Cinderella					Everyday				
		Story Grammar	Topic Coherence	Local Coherence	Reference chains	PAS	Story Grammar	Topic Coherence	Local Coherence	Reference chains	PAS
Moderate aphasia (n=7)	*Mean (SD)*	*3.14 (1.35)*	*39.14 (21.15)*	*4.69 (.28)*	*2.71 (1.98)*	*.79 (.22)*	*2.57 (.68)*	*51.48 (12.73)*	*4.61(.21)*	*1.5 (.78)*	*83 (.08)*
	Min - Max	2 – 5	0 - 55.56	4.34 - 5	1 – 6	.33 – 1	2 – 3.92	37.22 - 69	4.3 – 4.99	1.08 – 3.23	.72 - .95
Mild aphasia (n=10)	*Mean (SD)*	*3.7 (1.15)*	*46.37 (19.44)*	*4.73 (.17)*	*3.4 (2.37)*	*.84 (.14)*	*3.05 (.82)*	*62.06 (10.78)*	*4.62 (.24)*	*1.93 (1.02)*	*.87 (.14)*
	Min - Max	2 – 5	0 - 66.67	4.37 - 5	0 – 7	.63 - 1.09	2.38 - 5.15	40.33 - 73.81	4.26 - 4.9	.58 - 3.31	.85 -1.01
d effect size		*.45*	*.36*	*.17*	*.32*	*.28*	*.64*	*.9*	*.04*	*.48*	*.36*


Input trigger
Semantic memory

Episodic memory
Frame/ schema generation
Insertion of semantic information
Integration of semantic information
Top down and bottom up processing
Selection and topicalisation of information
Pragmatic judgments and requirements
Generation selection and chunking of propositions
Linguistic formulation
Articulation

Figure 1. Discourse production model from Sherratt (2007)

Appendix 1: Elicitation

Cinderella (Saffran et al. 1989; Webster, Franklin, and Howard 2007).

Participants were told they would be telling the story of Cinderella, and were offered the opportunity to familiarise themselves with the story, using a wordless picture book. If participants chose to familiarise themselves using the book, it was removed from view before testing. Then, participants were asked ‘Can you tell me the story of Cinderella?’.

Picture descriptions

Participants were presented with two black and white composite pictures, and asked: ‘Can you tell me about this picture?’. The picture was left in view whilst the participant was producing the discourse. The pictures were the ‘picnic’ scene from the Western Aphasia Battery-Revised (Kertesz 2007), and the ‘Cookie Theft’ (Goodglass and Kaplan 2001).

Procedural discourse (Cocks, Hird, Kirsner 2007; Pritchard, Cocks, Dipper and Morgan 2015)

Participants were asked to give instructions on how to carry out procedures, as though they were telling someone who had never completed the task before. Discourses were then elicited using the prompts ‘Can you tell me how you’d change the wheel of a car/ wrap a box in paper for a present?’. If a participant said they did not know, they were encouraged to describe as much as they knew of the procedure.

Personal discourses from the Autobiographical Memory Interview- AMI (Kopelman et al. 1990)

Personal discourses were elicited during the *AMI*, using the specific incident schedule from the published assessment, and prompts (table A1). Where participants were not able to identify a specific incident, predetermined prompts from the published assessment were used to elicit a discourse, for example, asking the participant about their first memory when attempting to elicit a discourse about a preschool experience.

--------------------------------------table A1 about here--

Table A1:
Autobiographical Incident Schedule and prompts, reproduced from the Autobiographical Memory Interview (Kopelman et al. 1990)

Time period	Episodic incident	Prompts
Childhood	Before school	Your first memory? Involving a brother or a sister?
	Primary school	Involving a teacher? Involving a friend?
	Secondary school	Involving a teacher? Involving a friend?
Early adult life	College/ first job	Your first day at work or college? An incident with a friend?
	A wedding	An incident involving a guest at the wedding? An incident at the reception?
	Meeting someone new	Meeting someone in an interview? Meeting someone on holiday or at work?
Recent times	Present setting/ an outpatient visit to hospital	Involving the warden? Involving the daily care staff? Involving the psychologist?
	Visit from a relative or visitor in the last year	A visit by or to a relative? Involving some news about a relative?
	Holiday or journey within the past year	At the place you visited? Involving someone you met?