Title: Treatment for Improving Discourse in Aphasia: a systematic review and synthesis of the evidence base

Running head: Discourse Treatment Systematic Review

Authorship: Lucy Dipper[1], Jane Marshall[1], Mary Boyle[2], Nicola Botting[1], Deborah Hersh[3], Madeleine Pritchard[1], Madeline Cruice[1]

Declaration of interest: Authors have no conflicts of interest to declare.

Contact details:
Dr Lucy Dipper, Associate Professor, Ph: +44 020 7040 4658. L.T.Dipper@city.ac.uk

Dr Madeline Cruice, Associate Professor, Ph: +44 020 7040 8290. M.Cruice@city.ac.uk

Professor Jane Marshall, Ph: +44 020 7040 4668. J.Marshall@city.ac.uk

Professor Nicola Botting, Ph: +44 020 7040 8314. Nicola.Botting.1@city.ac.uk

Dr Madeleine Pritchard. Madeleine.Pritchard.1@city.ac.uk

[1]Division of Language and Communication Science
School of Health Sciences
City, University of London
Northampton Square
London EC1V 0HB
England, UK

Professor Mary Boyle
[2]Department of Communication Sciences and Disorders
Montclair State University
1 Normal Avenue
Montclair, New Jersey 07043
United States of America
Ph: 973 655 7356
boylem@montclair.edu

Associate Professor Deborah Hersh
[3]School of Medical and Health Sciences
Edith Cowan University
270 Joondalup Drive
Joondalup WA 6027
Australia
Ph: +61 8 6304 2563
d.hersh@ecu.edu.au

Abstract

Background: Improved discourse production is a priority for all key stakeholders in aphasia rehabilitation. A Cochrane review of randomised controlled trials (RCTs) for aphasia found speech and language therapy treatment to be effective for improving ability to communicate in everyday interaction. However, this large-scale review did not focus exclusively on treatment for discourse production and did not include other treatment research designs. Thus, the extent of the evidence base addressing discourse interventions is currently unclear.

Objective: The present study undertakes the first systematic review of research on treatment for discourse production in aphasia, appraises the quality of the evidence base; characterises the methods for measuring outcomes; and describes discourse treatment in terms of both content and efficacy.

Design: Scopus, Medline, and EmBase databases were searched, providing 334 records. Twenty-five studies (reporting on 127 participants) met inclusion criteria and were reviewed with the following research questions: What is the quality of the study designs used? How complete is the intervention reporting? What is the range, type and content of outcome measures used? What is the range, type and content of discourse treatments reported to date? Are discourse treatments efficacious?

Results. Seven of the 25 studies met the criteria for quality review, with 3 RCTs scoring moderately well and 3 (of 4) case studies scoring moderate-low. Most studies had adequate levels of completeness of treatment reporting, with 3 scoring highly. There were 514 different outcome measures reported across the 25 studies, with measures of words-in-discourse the most common. Studies were grouped into six treatment categories: 'word production in discourse', 'sentence production in discourse', 'discourse macrostructure', 'discourse scripts', 'multi-level', and 'no consensus'. Twenty-two studies reported post-treatment gains, most commonly noted in increased word production. Changes in sentence production and discourse macrostructure were present, but infrequently assessed.

Conclusions: Discourse treatment is an emerging field of research. Despite limitations in the evidence base, there are clear positive signs that discourse treatment is efficacious. There is emerging evidence for beneficial effects on word and sentence production in discourse, for improved discourse macrostructure, and for treatments working at multiple levels of language. To strengthen the evidence in this field and improve outcomes for people with aphasia, we need more discourse treatment research using an explicit theoretical rationale, high-quality study designs, more complete reporting, and agreed treatment and assessment methods.

Introduction

Optimal recovery from aphasia is a top-ten research priority for life after stroke (Pollock et al., 2012) and improvement in discourse production is a critical component of this recovery. Discourse is core in everyday communication (Bryant et al., 2016), and improved discourse production is an outcome prioritised by people with chronic aphasia (Wallace et al., 2017). Recent international clinical guidelines also recommend that people with aphasia should be offered discourse treatment (e.g. the Australian Aphasia Rehabilitation Best Practice Statement, 2014; the Canadian Stroke Best Practice Recommendations: Guidelines update 2015). Thus, it is clear that improved discourse production is important to all key stakeholders, however there is limited guidance for researchers and clinicians on how to achieve this for people with aphasia as there is currently no synthesis of the discourse treatment literature.

In aphasia research, discourse is most often defined as either as a unit of language of a particular size (bigger than a single simple clause) or a form of language used for a specific purpose (storytelling, expressing an opinion, describing a procedure); or a combination of these two (Armstrong, 2000; Halliday, 2004). Discourse is therefore implicit in the form and content of everyday communication and when impaired in aphasia, subsequently affects a speaker's activities and social participation (Davidson et al., 2003), although impact is not routinely considered in definitions. In this study, we adopt the definition used by others (Armstrong, 2000; Halliday, 2004; Ulatowska et al., 1983), wherein discourse is composed of multiple linguistic levels that is, a level of *macrostructural organisation*, a level of *utterance or sentence structure*, and a *lexical* level. Throughout this paper the term 'discourse' is used to mean language beyond the clause, and the term 'discourse macrostructure' is used to refer to one of the linguistic levels that comprise discourse.

Evidence-based practice is essential for effective and efficient treatment of the discourse difficulties experienced by people with aphasia, and systematic reviews provide a method for evaluating the scientific evidence for specific treatment approaches. The 2016 Cochrane review of speech and language therapy for aphasia (Brady et al., 2016) emphasised the central role of treatments to improve the ability to communicate in everyday interaction. The primary outcome chosen to indicate the effectiveness of treatment was functional communication. The ideal measure of functional communication was proposed as communicative activity in real world settings, but few studies achieved this. Instead, the review found a range of methods used including functional communication assessments, and discourse assessed by picture description in test batteries or sampled in specific tasks. Because the review combined studies for analyses and reporting according to the broader objective of whether they impacted on communication, it is not possible to tease out the specific effect of interventions that specifically improve discourse. Thus, the extent of the evidence base addressing discourse treatment for aphasia is currently unclear.

Level and quality of evidence

Understanding how to achieve improved discourse is strongly influenced by the quality and completeness of the treatment evidence base. The purpose of the current review was to systematically evaluate the evidence that exists for the effects of spoken discourse production treatment in aphasia. The review aimed to: (1) appraise the quality of the evidence base; (2) characterise the methods for measuring outcomes; and (3) describe discourse treatment in terms of both content and efficacy.

The *levels of evidence* hierarchy in health care research is used to classifying evidence for literature synthesis (see for example, Cochrane Consumer Network – Level of Evidence, n.d.). In such frameworks, systematic reviews of randomized controlled trials (RCTs) are at the top of the

hierarchy, with single-case studies placed at the bottom. The highest level of primary research evidence for intervention studies is considered to be the randomised controlled clinical trial, although this design is less frequent in the aphasia treatment literature than other designs (Brady et al., 2016; Cherney et al., 2013). More frequent are single-cases, case series and non-randomized group designs which have been controlled by other means (comparison/control group or untreated condition/items). These are research designs are often viewed as a lower level of evidence. However, carefully controlled single-cases and case series can also provide quality evidence. Oxford University's Centre for Evidence-Based Medicine (OCEBM n.d.) propose that either single case randomized trials or systematic reviews of randomized trials provide level 1 evidence of efficacy to motivate clinical treatment choices (OCEBM Levels of Evidence Working Group. "The Oxford Levels of Evidence 2", 2011). Within these broad categories of experimental design, the specifics of a study's design should also be evaluated to appraise *quality*, where possible. There are published tools for appraising quality. The modified Physiotherapy Evidence Database (PEDro-P) scale is used for assessing quality of controlled group trials (Maher et al., 2003) and the Risk of Bias in N-of-1 Trials scale (RoBiNT) is used for for assessing quality of single-case designs (Tate et al., 2015). Other study designs, such as non-randomised between-group studies or within-group studies which control by other means cannot be appraised for *quality* using the same methods but can be appraised more broadly using hierarchies of evidence in terms of the *level* of evidence they provide, as described above. An additional step in evaluating treatment evidence is a consideration of the *comprehensiveness* of treatment reporting, which is essential for replication and implementation. Hoffman et al. (2014) provide a template checklist for RCTs to guide the reporting process so that interventions are described in enough detail to allow their replication. The checklist has also been used retrospectively to evaluate the completeness of treatment reporting (e.g. Brady et al., 2016).

Outcome measurement

A crucial component of treatment evidence is the measurement of outcome. Previous reviews of discourse assessment in aphasia reveal a large number of measures and methods have been used by researchers and there is no consensus on the best outcome measure(s) for discourse treatment (Armstrong, 2000; Bryant et al., 2016 ; Ellis et al., 2016; Linnik et al., 2015; Pritchard et al., 2017). Whilst this proliferation of discourse measures indicates a positive trend towards more functional and meaningful outcomes of linguistic treatment, the negative consequences are the variable psychometric robustness of new measures and a lack of agreement about the most representative outcomes (Dietz & Boyle, 2018). It is also unclear how outcomes at the different linguistic levels of words, sentences and discourse macrostructure relate to each other. Webster et al. (2015) reviewed the treatment literature for spoken production at these three linguistic levels and focused on the methods used to predict and measure generalisation across levels. They found a limited evidence base from which to decide *where* and *how* to best measure generalisation and subsequently argued for the need to develop a greater understanding of the relationship between change at different linguistic levels, using both theory and evidence. Thus, the existing context for considering the outcome measures in this review is a broad range of measures and methods and little consensus. Alongside this context is an emerging awareness from researchers that outcomes need to be better planned and intentionally targeted at expected change.

Discourse as a Complex Intervention

Discourse treatments are, by their nature, complex interventions (Craig et al., 2008) and there is no agreed definition about what constitutes a discourse treatment. A major challenge is the need to target multiple linguistic levels and there is limited discussion about how these levels relate to each other. A second challenge is that, despite such treatments aiming to change discourse, they use a variety of different treatment methods to do so. For example, a study might have discourse as its treatment goal, but adopt a word-level focus in treatment to achieve this aim, using a variety of

word-based treatment activities such as picture naming and lexical-semantic tasks to train words that are then practised in a discourse context. In the review mentioned above, Webster and colleagues (2015) grouped papers (with varying outcome goals) according to treatment type. They divided them broadly into single level or multi-level treatments, and then subdivided the single level studies according to the target of the treatment (word, sentence or discourse). Studies designed to improve discourse production appeared in more than one of these treatment groups (i.e. 'discourse level therapies'; 'word and connected speech level therapies'; 'sentence and discourse level therapies'; and 'word, sentence and discourse level therapies') which suggests that treatment for discourse is not a homogenous group. Consideration of two multi-level treatment studies reviewed by Webster et al. (2015) illustrates some of the complexity of discourse treatment. Carragher et al. (2015) targeted discourse production in the context of a video story retelling task. Treatment focussed on the training of key strategies, which were individualised to each participant based on their communication needs. Examples include, moving on when experiencing word-finding difficulties and planning the story in small chunks of information. Additionally, communication partners received conversation coaching to enable them to develop strategies to check and clarify information. Whitworth et al. (2015) targeted the production of a range of everyday discourse genres such as recount, procedure and exposition. Treatment focussed on the description of picture sequences, using visual frameworks for sentence and story structure. These two studies are similar in terms of their discourse goals and both encompassed multiple treatment levels (words, sentences, and discourse macrostructure). They differ however in terms of *how* the treatment goals were addressed, with the use of strategies in Carragher et al. (2015) and visual frameworks in Whitworth et al. (2015). In the last decade, there has been increased interest in *how* treatments work which are relevant for consideration in our review. Although Baker (2012) writes about optimising treatment intensity which is not a focus of this review, she highlights the central importance of *how* intervention goals are addressed which she defines as *therapeutic inputs* (e.g. modelling, cueing, or feedback from the clinician) and *client acts* (production or practice of a specific

skill). In the description of their treatment, Carragher et al. (2015) and Whitworth et al. (2015) mention different therapeutic inputs. Carragher and colleagues highlight facilitation, prompting, visual record-keeping, modelling; Whitworth and colleagues mention written record-keeping and feedback. This is a further aspect of complexity in the discourse treatment evidence base under review. We draw on Baker's definitions to enable us to identify the linguistic level(s) specifically targeted by therapeutic activity in each study, and thereby categorise this complex evidence base into treatment groups.

The current study

In summary, discourse interventions for people with aphasia exist but are reported within a wider body of treatment literature i.e. reviews by Brady et al. (2016) and Webster et al. (2015). A synthesis of discourse-specific aphasia treatments is needed to consolidate research to date and inform future research and practice. The objective of this review is to examine treatment research studies which explicitly target discourse in treatment as well as measure discourse as an outcome from treatment. The three overall aims for this review and synthesis yield the following research questions:

Aim 1: to appraise the quality of the evidence base

RQ 1: What is the quality of the study designs used?

RQ 2: How complete is the intervention reporting?

Aim 2: to characterise the way outcomes are measured

RQ 3: What is the range, type and content of outcome measures used?

Aim 3: to describe discourse treatment

RQ 4: What is the range, type and content of discourse treatments reported to date?

RQ 5: Are discourse treatments efficacious?

The methods used to address these questions are outlined in the next section.

Methods

Search string

Scopus, Medline, and EmBase databases were searched twice, on 25/05/2018 and 18/07/2018. In the first search the terms ['discourse' or 'narrative' or 'story' or 'storytelling'] and ['intervention' or 'treatment' or 'therapy'] and ['aphasia' or 'dysphasia'] in the TITLE-ABS-KEY fields were used. In the second search the terms ['connected speech'] and ['intervention' or 'treatment' or 'therapy'] and ['aphasia' or 'dysphasia'] in the TITLE-ABS-KEY were used. Table 1 details the exclusion and inclusion criteria, and lists the papers included in the review. Database searches and hand searching yielded a total of 334 records for initial consideration. Of these, 309 were discarded because they did not meet the criteria. Therefore, 25 studies were included in the review (see Figure 1), and report on 127 participants with aphasia (see supplemental file 1 for more information on participant demographics).

[Insert Figure 1 and Table 1 about here]

Method for RQ 1 (What is the quality of the study designs used?)

Where studies met the criteria for doing so, we used two critical appraisal tools for evaluating rehabilitation research design. The PEDro-P (Maher et al., 2003) is a validated tool for evaluating the quality of controlled group trials. To assess whether a study was a randomised rather than non-randomised control trial and could therefore be assessed using the PEDro-P scale (PsychBITE, 2018), authors' descriptions of the study design in the abstract and methods sections were assessed against

the definition[1] from NICE (2018). Consequently, those studies reporting within-subjects, wait-control, or non-randomised between-group designs were not assessed using PEDro-P. The RoBiNT (Tate et al., 2015) is validated for evaluating the quality of single-case designs. The RoBiNT is a 15-item scale using a 0-2 rating criteria, where 2 points are awarded if the design meets the stringent criteria; 1 point awarded if the design had been acceptable in the past; and 0 points if the study failed to meet minimum standards. Both scales evaluate quality with reference to a predetermined list of design elements which relate to the internal validity of the study (Appendix 1 for details).

Seven of the 25 studies in the review met the criteria for appraisal using these tools. The other 18 studies did not have study designs appropriate for *quality*-review using these tools, although the *level* of the evidence they provided can be evaluated in terms of established evidence hierarchies which place them below RCTs. Inter-rater reliability for both PEDro-P and RoBiNT scoring was assessed. Two judges (LD and MP) scored 100% of the studies that were suitable for evaluation with the PEDro-P scale (n=3) and RoBiNT scale (n=4). For the PEDRO-P scale, agreement was 91% (30 out of 33 potential agreements: 3 studies, each rated on 11 scale items). For the RoBiNT scale, agreement was 72% (43 out of 60 potential agreements: 4 studies each rated on 15 scale items). Non-agreements[2] were resolved through discussion, re-checking the manuals, and consulting an experienced user of the scales. Scores from PEDro-P and ROBIN-T scales are summarised and reported descriptively.

[1] *A study in which a number of similar people are randomly assigned to 2 (or more) groups to test a specific drug, treatment or other intervention. One group (the experimental group) has the intervention being tested, the other (the comparison or control group) has an alternative intervention, a dummy intervention (placebo) or no intervention at all. The groups are followed up to see how effective the experimental intervention was. Outcomes are measured at specific times and any difference in response between the groups is assessed statistically. This method is also used to reduce bias*. (NICE, 2018).

[2] *Non-agreements for the ROBIN-T were for sampling of behaviour (n=4 studies); dependent variable (n=3); design with control (n=3); generalisation (n=2); raw data record (n=2); and a single non-agreement each for the items setting, interrater reliability, and data analysis. The non-agreements were generally due to different interpretations of scoring criteria, or differences interpreting the information reported in the studies. For example, for item 'sampling of behaviour', the judges differed in their interpretation of what constituted a data point; and for the item 'dependent variables', the judges differed in their interpretation of what constituted an operational definition of a target behaviour, and clarity/precision of method of measurement.*

Method for RQ 2 (How complete is the intervention reporting?)

The completeness of intervention reporting was assessed for all studies using the 12 item Template for Intervention Description and Replication (TIDieR) (Hoffman et al., 2014). Whilst intended for description of treatments provided in RCTs, this has been applied more broadly to treatment studies of all designs. Using this checklist, two authors (MC and MP) independently extracted from the studies the information required for the TIDieR. Disagreements were discussed to reach consensus, and then each study was given a consensus score (out of 12) for the completeness of reporting. See Appendix 2 for full details of TIDieR items and definitions.

Method for RQ 3 (What is the range, type and content of outcome measures used?)

All outcome measures within the studies were extracted into an Excel spreadsheet by one author (MP) and checked by two further authors (MC and LD). Following Pritchard et al. (2017), we counted each individual 'incident' of measure use (i.e. every use of a measure that returned a score). For example, where a study counted treated nouns and untreated nouns and returned a separate score for each, this was counted as two measure incidents. Similarly, where studies used a measure and reported it both as count and as a percentage, this was counted as two measure incidents. The extracted outcome measure incidents were then categorised descriptively, based on the type of measure (standardised[3] or not) and the linguistic level they measured (word, sentence, macrostructure). A further distinction was needed to categorise whether words or sentences were assessed in a specific bespoke task outside of discourse (e.g. naming or sentence production with picture stimuli) or whether a score was extracted from a discourse task (e.g. recount). These distinctions yielded the following categories of measures: *bespoke* single word production tasks

[3]The term 'standardised' here indicates that a test or test battery is commercially available, has a standard form of administration and has published normative and/or clinical data available.

(e.g., a naming test created for a specific set of treated words); *standardised* single word production tasks (e.g., The Boston Naming Test, Kaplan, Goodglass & Weintraub, 1983); single word measures (word or morphology production, or errors) *extracted* from discourse (e.g., number of incorrect morphemes produced in a discourse); *bespoke* sentence production tasks; *standardised* sentence production tasks; sentence measures *extracted* from discourse; discourse macrostructure tasks; and Aphasia Batteries.

Method for RQ 4 (What is the range, type and content of discourse treatments reported to date?)

Discourse is multifaceted, and assessment and treatment of it can cover a range of aspects, from single words, to the information a speaker communicates. This review necessarily captures (through inclusion criteria) only those treatments that included discourse work, i.e. language work in a discourse context or direct work on discourse macrostructure. Nonetheless, each study has a specific focus for the main therapeutic inputs (Baker, 2012). For example, a study with the goal of improving sentence production in discourse might have the majority of therapeutic *activity* at sentence level (tasks involving sentences being produced in isolation, rather than in a discourse) but will also have additional practice of these sentence level targets in a *discourse context*. Such studies will have been included in our review. Similarly, treatments with the goal of improving word retrieval in discourse are included in the review if they involve at least some practice of word production in discourse, even though their main therapeutic activity may have occurred in single word treatment tasks. In summary, all treatments in this review focussed on therapeutic work aimed at improving discourse production but not all included work on discourse macrostructure.

To enable comparison of studies across all three review aims, studies were categorised into treatment groups based on the linguistic level(s) specifically targeted during treatment. To achieve this, we agreed an operational definition of therapeutic activity, as follows: an explicit statement

that some form of therapeutic activity such as cueing, correction, feedback or scaffolding was provided by the clinician for a particular activity. Our operational definition is in line with Baker's (2012) definition of "the ingredients that make up an intervention" (p483). Once identified as a therapeutic activity, the linguistic level targeted by the activity could then be coded and the study thereby categorised into a treatment group (i.e. the study could be grouped with other studies which also targeted the same linguistic level(s)) – see Table 2 for resulting treatment groups. Three members of the research team (MP, JM, and MB) categorised all studies independently using the full contents of the published studies (including appendices and online supplemental files). There was complete agreement (3/3 coders) for 16 studies (64%) on first categorisation (Antonucci, 2009; Carlomagno & Losanno, 1991; Cherney, 2010; Cherney & Halper, 2008; Cherney et al., 2008; Cherney, 2010; Falconer & Antonucci, 2012; Goral & Kempler, 2009; Greenwood et al., 2010; Hickin et al., 2015; Lee et al., 2009; Murray et al., 2007; Penn & Beecham, 1992; Wambaugh et al.,2013; Whitworth, 2010; Whitworth et al., 2015). There was also majority agreement (i.e. 2/3 coders) for a further 5 studies (20%) on first categorisation (Carragher et al., 2015; Hoover et al., 2015; Marini et al., 2007; Milman et al., 2014; Nickels et al., 2016). To reach further consensus we discussed which activities described in each study were therapeutic, using the operational definition outlined above. Disagreements were re-considered on this basis, and consensus thereby reached for a further 2 studies (Gordon, 2007; Osiejuk, 1991) which resulted in final agreement for 23/25 studies (92%). No consensus was reached about which treatment group the final two studies should be categorised into for Dietz et al. (2018) and McCall et al. (2009). This was not necessarily due to lack of clarity in the reporting, but because the team could not agree on whether the focus of therapeutic activity was single or multi-level. Therefore, these two studies have been termed 'no-consensus'.

[Insert Table 2 here]

Method for RQ 5 (Are discourse treatments efficacious?)

Information about efficacy was extracted from each study into an Excel sheet by one author (MP) and checked by two further authors (MC and LD). Studies were so varied in terms of research design and outcome measures which meant that quantitative analysis and synthesis of efficacy was not possible. Instead the extracted data regarding treatment efficacy was summarised descriptively.

Results

RQ1: What is the quality of the evidence base?

Three of the 25 studies met the criteria for evaluation with the PEDro-P, 4 met the criteria for evaluation with the RoBiNT, and eighteen did not have a study design suitable for evaluation using either scale (7 were single-case reports with no control; 6 were cases series; 3 were group studies with no control, and 2 were non-randomised group studies with control – one using a repeated measures design and the other delayed treatment) .

All three RCTs received moderately good scores on the 10-point PEDro-P scale. Two trials (Cherney, 2010a; Dietz et al., 2018) scored 6/10, and the remaining trial (Whitworth et al., 2015) scored 7/10. All three studies lost points on the following criteria: concealed allocation, blinding of subjects, and blinding of treating therapists. Whitworth et al. (2015) reported blinding of outcome assessors, accounting for its higher score. Four studies were assessed with the ROBIN-T scale (Goral et al.2009; Greenwood et al., 2010; Hickin et al., 2015; Murray et al., 2007). These case studies received moderate-low scores on the 30-point RoBiNT scale (scores ranged from 8-13 out of 30). In terms of internal validity, no study received points for randomization, blinding of participants, blinding of assessors, or treatment adherence. In terms of external validity, no study received points for replication, reflecting the fact that treatment of discourse is still in an early phase of investigation. Results are summarized in Table 3.

[Insert Table 3 here]

RQ2: How complete is the intervention reporting?

The level of reporting completeness for the interventions included in this review varied (see Table 4). Studies reported on average 6.7 items (range 3-12) with the majority of studies reporting 5 or 6 items. Three studies scored highly at 11 or 12 and were the scripting studies using technology. Considering TIDierR items, interventions are most completely reported for item 8 (dosage, duration, and intensity in 25/25 studies) and item 4 (treatment procedures in 24/25). See supplemental file 2 for further information on treatment dose and duration. Other well reported items were items 3 (treatment materials) and 6 (mode of delivery) in 22/25 studies each. Intervention names and rationales, theoretical underpinning or goals of essential elements (or the intervention as a whole) were reasonably reported (17/25 studies). Remaining aspects of intervention provider qualifications, location of delivery, and personalisation and fidelity were minimally reported.

[Insert Table 4 here]

RQ3: What is the range, type and content of outcome measures used?

All studies used multiple outcome measures, with 514[4] incidents of measure use (pre and post treatment) extracted from the 25 studies (mean=21 measure incidents/per study; SD= 16.81; range= 2-70). The study reporting 70 measure incidents was Murray et al. (2007) in which 4 discourse samples (2 written, 2 spoken) were scored and individually reported for the following: (1) total number of words and (2) words per minute; (3) total number of CIUs and (4) CIUs/minute; (5) proportion of CIUs; (6) ratio of open to closed class words; (7) number of utterances; (8) mean

[4] One study (Penn & Beecham 1992) had a multilingual participant for whom scores were reported in 4 languages. We have only reported the counts for the measure incidents reported for English, so as not to distort the findings.

length of utterance (MLU); proportions of (9) grammatically complete and (10) grammatically complex utterances; and (11) number of substantive verb productions. Additionally, 6 of these scores were also reported for the mean across both written discourses; 10 reported for the mean across both spoken discourses; and 10 reported for the mean across all four discourses.

Overall, discourse treatments have been evaluated in a range of ways. This included via a range of functional communication assessments: *Communication Activities of Daily Living* (CADL: Holland, 1980); *The Communication Disability Profile* (CDP: Swinburn & Byng, 2006); the *American Speech-Language Hearing Association Functional Assessment of Communication Skills for Adults* (ASHA FACS: Frattali et al., 1995); and the *Communicative Effectiveness Index* (CETI: Lomas et al., 1989). Other tests included impact assessments namely *the Assessment for Living with Aphasia* (ALA: Kagan et al., 2011) and the *Quality of Communication Life Scale* (ASHA QCL: Paul et al., 2004), and aphasia batteries namely *the Western Aphasia Battery* (WAB: Kertesz, 1992). Standardised tests or tasks of naming and sentence production were also noted: the *Boston Naming Test* (BNT: Kaplan et al., 1983), the *Philadelphia Naming Test* (PNT: Walker & Schwartz, 2012), and the sentence production test by Caplan and Hanna (1998). Bespoke tasks targeting word and sentence production (i.e. tasks such as picture naming, specifically designed for treated and untreated items) were also frequently reported. Finally, treatment was also evaluated using scores for word and sentence production extracted from the production of a full discourse, and from tasks in which the discourse macrostructure was assessed (e.g. using counts of Story Grammar elements). See Figure 2 for a visual summary of types of measure used by each treatment group. Although there were only a few studies in each treatment group, there was a relatively large number and type of incidents of outcome measures employed within each group (see Table 5 for detail).

[Insert Figure 2 and Table 5 about here]

Unsurprisingly, for treatments targeting words in discourse, the most commonly used outcome measures assessed progress on single words. The method of assessment varied however, with some studies using naming tasks (such as picture naming), and others counting either single words, morphology, or single word errors from within discourse samples (such as counting the number of treated nouns produced without error in a discourse). Two studies in the word production in discourse treatment category (Antonucci, 2009; Falconer & Antonucci, 2012) also used the WAB but did not report on the event description separately. The treatments targeting sentence production in discourse, discourse macrostructure, and discourse scripts did *not* use naming tools, but rather extracted single words and sentences from discourse, and also used aphasia batteries. As expected, treatments targeting multiple levels used the widest range of outcome measures across all linguistic levels, and included both standardised tests and information extracted from discourse. In 4 multi-level studies both these methods were used (Hoover et al., 2015; Marini et al., 2007; Milman et al., 2014; Whitworth et al., 2015). The multi-level treatment category included studies that also reported measures of discourse macrostructure, and this was the only treatment group to include studies that did so, except for one study in the discourse macrostructure treatment category (Osiejuk, 1991).

RQ4: What is the range, type and content of discourse treatments reported to date?

Studies were grouped based on the linguistic level specifically targeted for change, at which the *main therapeutic activity* occurred. This resulted in 5 word production in discourse treatment studies; 5 sentence production in discourse treatment studies; 2 discourse macrostructure treatment studies; 3 discourse script treatment studies; 5 studies reporting treatment at 2 levels of language; 3 studies reporting treatment at 3 levels of language; and 2 final studies where we could not reach consensus on which linguistic levels were therapeutically targeted.

Table 6 summarises the key treatment methods used in each study. These methods were extracted from the studies during completion of each TIDieR checklist in response to items 3 and 4 (materials and procedures). As Table 6 indicates, there were some commonalities in methods across studies, both within and between groups. The following sections synthesises the characteristics of each treatment group.

[Insert Table 6 about here]

Word production in discourse treatment studies

In the 'word production in discourse' group, *the treatment targets* were primarily nouns, which were treated using picture naming, repetition, and cueing based activities; or by semantic treatment inspired by semantic feature analysis (SFA); and/or target-word production in response to focused discussion in group treatment. The *discourse contexts* in which treatment targets (words) were practised varied from study to study and included PACE-like group activities, story retelling with clinician cueing (of words), and conversation or group discussion to practice treated words.

Sentence production in discourse treatment studies

In the 'sentence production in discourse' group, *the treatment targets* included spoken sentences which were either read aloud, repeated, or generated in response to a picture or in a communicative task. Some studies also used written sentences. As the focus of this review is spoken language, these studies were only be included as the written items were used in addition to the spoken items. Some studies additionally included a meta-linguistic task such as identifying thematic roles (such as agent and theme) in a written sentence or rearranging sentence fragments (words or phrases) to create a sentence. *The discourse contexts* in which treatment items (sentences) were practised again varied

across studies. They generally involved creating a discourse using treated sentences by describing pictures, playing games, making phone calls, generating stories, recounting episodes, and having conversations.

Discourse script treatment studies

In the 'discourse script' studies, *the treatment targets* were personalised *discourse* scripts to be read aloud and repeated. Therapy was delivered using AphasiaScripts™, a software program with a virtual therapist who provides the therapy stimuli through two modalities: audio (a recorded version of the script) and visual (watching the therapist's oral motor movements, and seeing the written sentences on the screen, synced with the audio).

Discourse macrostructure treatment studies

In the 'discourse macrostructure' group, both studies reported discourse macrostructure *treatment targets*, i.e. treatment activities designed to improve the macrostructure of the following types of discourse*:* narrative discourse, procedural discourse (the description of a procedure, such as making a cup of tea, or changing the wheel on a car), life events, and story retell from video clips. The methods in each study were different but involved discourse production exercises using pictures, oral stories, or videos as stimuli. In both studies, the discourse to be produced was structured or supported with either templates about what information should be included or with particular strategies (e.g. identifying the main referent or creating a visual record). Cues and prompts from the clinician addressed discourse macrostructure in general, and also addressed strategies for signalling discourse macrostructure (e.g. the use of discourse connectives such as 'next').

Multi-level treatment studies

The 'multi-level' studies, by definition, treated more than one linguistic level. In the evidence base we found both a group of studies which reported therapeutic activities at two levels and a smaller group that reported therapeutic activity at three levels of language. All multi-level studies reported therapeutic activity at the discourse macrostructure level, but they varied in terms of what was targeted at other levels. The two-level studies combined *treatment targets* at discourse macrostructure and word level (Carlomagno et al., 1991; Milman et al., 2014; Penn & Beecham, 1992) or at discourse macrostructure and sentence level (Hickin et al., 2015; Marini et al., 2007). Treatment methods included Promoting Aphasics' Communication Effectiveness (PACE: Davis & Wilcox, 1981); the Helm Elicited Language Program for Syntax Stimulation (HELPSS: Helm-Estabrooks, 1981), and Response Elaboration Treatment (RET: Kearns, 1986), as well as discourse production exercises. The *discourse contexts* for the two-level studies included personalized discourses (topics of interest such as 'what did you do this morning?'), picture and cartoon descriptions, riddle play using famous people, and story completion. The three-level studies each included therapeutic activity at word, sentence and discourse macrostructure levels. *Treatment targets* were verbs (word level), verb-argument structures (sentence level), and either story grammar structure or discussion topics (discourse macrostructure level). *Discourse contexts* included personalized discourses (event recall, event planning, providing opinions), picture descriptions, and discourse generated using topic headings.

Studies with 'no consensus' on treatment activity

The final two studies in our review could not be classified into the treatment categories. Both studies used technology to support therapeutic activity: a Dynavox VMAX AAC device, with treated personal narratives programmed into the device (Dietz et al., 2018), and SentenceShaper software in which treated sentences were programmed (McCall et al., 2009). In both cases there was the opportunity for people to work on word retrieval, sentence production, and discourse production

using the technology. However, it was not clear from the treatment reporting whether there was therapeutic activity (according to our operational definition) at each of these levels.

RQ5: Is there evidence that discourse treatments are efficacious?

In this section, the data about treatment effects extracted from the studies are synthesised within treatment group. For the fuller detail about treatment effects for each study, please see Appendix 3.

Word production in discourse treatment studies

As Table 7 shows, the *word production in discourse studies* group includes data from 12 participants, reported across five studies (two studies recruited additional participants who did not complete the intervention: Antonucci, 2009 and Nickels et al., 2006). Changes in discourse production as a result of therapy were reported in three studies (Antonucci, 2009; Falconer & Antonucci, 2009; Greenwood et al., 2010), with 7 participants demonstrating improvements. In all cases, discourse benefits were evident in single word or morphology production, and for 6 these were maintained at follow up. Two studies (Antonucci 2009; Falconer & Antonucci, 2009) used effect sizes to evaluate the significance of the change, while in the third (Greenwood et al., 2010) discourse data were purely descriptive. Discourse changes at the sentence or macrostructure levels were not explored in any of the studies. With respect to non-discourse measures, all studies employed bespoke or standardised naming tools. Four participants demonstrated changes on these tools that were attributed to therapy (Falconer & Antonucci, 2009 (P4 only); Gordon, 2007; Greenwood et al., 2010). Naming gains were maintained in Gordon (2007) and Greenwood et al. (2010), and not assessed in Falconer and Antonucci (2009). Aphasia batteries were employed in three studies (Antonucci 2009; Falconer & Antonucci, 2009; Gordon 2007). Only Gordon (2007) reported substantial change on these batteries that exceeded the clinically significant level of 5 points; this was evident for both participants and

maintained at follow up. One study (Greenwood et al., 2010) used the CDP (Swinburn & Byng, 2006) to explore therapy impacts on activity, participation and emotional consequences, with positive findings.

[Insert Table 7 here]

Sentence production in discourse treatment studies

Five studies evaluated *sentence production in discourse* treatment, involving 30 participants (Table 8). All five studies report treatment-induced changes in discourse production at the word or morphological level. In four studies changes were evaluated statistically (Cherney 2010 a&b; Goral & Kempler, 2009; Wambaugh et al., 2013); while Murray et al. (2007) report descriptive data only. Benefits were influenced by severity and genre. In Cherney (2010b) benefits occurred for those with mild and moderate aphasia (N = 19), but not severe aphasia (N = 6). In Wambaugh et al. (2013) changes were detected in procedural but not personal discourse. One study (Murray et al., 2007) probed for cross-modality generalisation, with the finding that therapy conducted in writing benefited spoken discourse. In three studies, lexical gains in discourse were evident at follow up (Goral & Kempler, 2009; Murray et al., 2007; Wambaugh et al., 2013), with maintenance not assessed in the Cherney studies (2010a&b). Despite the intentional sentence focus of therapy, only one study probed for changes in sentence production extracted from discourse, with no clear benefits (Murray et al., 2007). None of the studies explored change in macrostructure within discourse. With respect to non-discourse measures, three studies employed aphasia batteries, with two showing change (Cherney 2010 a&b). In Cherney (2010b) group analysis of change did not exceed the clinically significant five points, however there were individual participants who achieved an increase of five or more points. The participant studied by Murray et al. (2007) demonstrated improvements on a bespoke sentence production task, which were maintained at follow up. Goral et

al. (2009) employed a novel social-communicative rating scale in their single case study and reported post therapy gains (maintenance was not assessed).

[Insert Table 8 here]

Discourse script treatment studies

As Table 9 shows, there were three studies, involving 23 participants, which evaluated *discourse script therapies*. All three studies reported discourse benefits at the level of single word or morphology, as demonstrated in the production of practiced scripts. In one study (Cherney et al., 2008) improvements were made by two (out of three) participants. In the others, gains were either achieved by all participants (Cherney and Halper 2008) or were demonstrated across the mean group score (Lee et al., 2009). The reported changes in discourse were not evaluated statistically in any of the studies and maintenance was not assessed. None of the studies evaluated discourse production at the sentence or macrostructure levels. Two studies employed aphasia batteries in addition to the discourse measures (Cherney et al., 2008; Cherney and Halper 2008), with patterns of change that were inconsistent across participants and measures.

[Insert Table 9 here]

Discourse macrostructure treatment studies

Two studies evaluated *discourse macrostructure therapy* involving five participants (Table 10). Both studies reported discourse improvements at the level of word production, achieved by all participants, although in one these were more evident in simple than complex narratives (Carragher et al., 2015). Both studies also reported changes in macro structure, e.g. affecting narrative sequence (Carragher et al., 2015). Only Osiejuk (1991) explored sentence level changes following therapy, and the participant produced more clauses but no fewer grammatical errors. Neither study

employed statistics nor assessed maintenance. Reporting in Osiejuk (1991) was highly descriptive, with some reported changes lacking corroborating data. An aphasia battery (the Boston Diagnostic Aphasia Examination, Goodglass and Kaplan, 1972) was administered in Osiejuk (1991) with reported changes in auditory comprehension and oral expression.

[Insert Table 10 here]

Multi-level treatment studies

Five studies, involving 16 participants, employed *multi-level treatments* which targeted *2 linguistic levels* (Table 11). Four studies assessed and reported change in the production of words in discourse (Carlomagno et al., 1991; Hickin et al., 2015; Marini et al., 2007; Milman et al., 2014). In two studies (Carlomagno et al., 1991; Marini et al., 2007) improvements were demonstrated in group means. In the other two studies, three (out of four) participants improved. Maintenance was only assessed in one study, and was reported to be good (Marini et al., 2007). These word level changes in discourse were evaluated statistically in only two studies (Carlomagno et al., 1991; Marini et al., 2007). Two studies, involving 4 participants, explored sentence level changes in discourse (Hickin et al., 2015; Marini et al., 2007). Both reported significant benefits post therapy, but with no information about maintenance. Macrostructure changes were explored in two studies (Hickin et al., 2015; Penn & Beecham, 1992). The participant in Hickin et al. (2015) made no changes on measures of coherence and story grammar. The individual studied by Penn and Beecham (1992) improved in narrative organisation. Maintenance was not assessed, and reporting was descriptive. Non-discourse measures included aphasia batteries (Hickin et al., 2015; Marini et al., 2007; Milman et al., 2014), standardised and bespoke naming tools (Milman et al., 2014), and bespoke sentence production tasks (Milman et al., 2014). Gains were most consistently reported in Milman et al. (2014) for the two participants who completed intervention. Other reported changes affected quality of life (Hickin

et al., 2015), functional communication (Marini et al., 2007) and compensatory strategy use (Penn & Beecham, 1992).

[Insert Table 11 here]

Three studies, involving 28 participants, evaluated *multi-level treatment at 3 linguistic levels* (Table 12). All studies assessed single word production in discourse. Significant changes were reported for one participant (out of two) in Whitworth (2010) and for the eight participants who received NARNIA therapy in Whitworth et al. (2015). Post therapy gains were maintained at follow up. Sentence production in discourse was also assessed in all studies, and all reported significant and maintained gains on at least some of the assessed parameters. Macrostructural features of discourse were assessed in two studies, with positive findings (Whitworth 2010; Whitworth et al., 2015). Maintenance of macrostructure gains was not achieved by the NARNIA therapy group in Whitworth et al. (2015). Turning to non-discourse measures, the 12 participants in Hoover et al. (2015) achieved significant and maintained gains on standardised naming tools. There were also gains on a standardised sentence production tool and on aphasia batteries, although maintenance was less clear, or in the case of the batteries, not assessed. Gains on standardised naming tools were also reported in Whitworth et al. (2015), but not on a standardised sentence production tool.

[insert Table 12 here]

Studies with 'no consensus' on treatment activity

For the final two studies (*no consensus* studies, which were not categorised into the other treatment groups), 13 participants were involved (Table 13). Both studies evaluated change in single word and/or morphology production in discourse, with significant benefits following therapy. Sentence production in discourse also improved in both studies. Macrostructure change was only assessed by

Dietz et al. (2018). The AAC group (the 'treated' group) demonstrated positive change with a reduction in linguistic non-fluencies known as *mazes* (filled pauses, repetitions, revisions, and abandoned utterances). Dietz et al. (2018) also administered an aphasia battery, in which the AAC group improved. Maintenance of change was assessed in neither study.

[Insert Table 13 here]

Overview of review findings

In summary, 25 studies reporting on 127 participants with aphasia were included in this review. Seven of the studies met the criteria for quality review (using PEDro-P and RoBiNT), with three RCTs scoring moderately well and three of the four controlled case studies scoring moderately. In terms of completeness of treatment reporting (using the TIDieR), the majority of studies scored moderately, with three studies scoring highly. Five hundred and fourteen (514) outcomes were measured across the 25 studies. Outcomes comprised discourse measures; clinical assessments of naming, sentence production, language functioning (batteries), functional communication, and impact; and bespoke naming and sentence production tools. Studies could be grouped into six treatment groups (word production in discourse, sentence production in discourse, discourse macrostructure, discourse script, multi-level 2 and multi-level 3), and consensus was reached on this for all but two of the 25 studies. Twenty two of the 25 reviewed studies (88%) reported post-treatment gains, with the most commonly reported gains (21 studies) relating to the production of words. Changes in sentence production and discourse macrostructure were less frequently assessed and reported.

Discussion

This is the first systematic review to fully address the evidence for improving discourse in people with aphasia. It highlights that whilst discourse is a key priority for a range of stakeholders in aphasia rehabilitation, the current evidence base is small and difficult to interpret. There is a lack of consensus across studies about what discourse treatment is. Our findings overall point to an emerging evidence base, consisting of study designs of variable quality and lack of detail in treatment reporting. The review also found notable exceptions. There are a few stronger, well-described studies in the evidence base which should be used to move the field forward (see, for example, Hoover et al., 2015; Whitworth et al., 2015). The discourse treatment evidence base is also characterised by diversity in both treatment and assessment methods. However, there is some unity across studies with the most common measure being single words extracted from discourse. These characteristics make it difficult to synthesise and compare across studies and limits interpretation of the results. These issues are unpacked in the following sections, in which the findings for each research question are interpreted.

Level and quality of evidence

In appraising the discourse treatment evidence base, we found seven studies with research designs that could yield high level evidence of treatment efficacy and that could be quality-rated using the standardised rating scales. The randomised clinical trials explored: (1) the effects on discourse production of AAC device use, with improvements at all three linguistic levels (word, sentence and discourse macrostructure) (Dietz et al., 2018); (2) therapeutic activity at the sentence level, with improvements on words in discourse (Cherney, 2010a); and (3) therapeutic activity at multiple levels of language, with improvements at all three levels (Whitworth et al., 2015). These RCTs scored moderately well on quality (60-70%). They scored adequately on completeness of reporting (42-58%), but lost points for lack of concealed allocation, blinding of subjects, blinding of treating therapists, and in some cases lack of blinding of outcome assessors. Despite these minor limitations,

these studies show promise for moving the field forward, and the multi-level approach to treatment is particularly promising given the improvements shown for words, sentences and discourse macrostructure.

Well-controlled case studies have the potential to yield high-quality evidence for clinical decision-making. The four case studies explored: (1) the effects on discourse of therapeutic activity at the word level, reporting word gains (Greenwood et al., 2010); (2) therapeutic activity at the sentence level (Goral & Kempler, 2009; Murray et al., 2007), reporting words gains in words in both, and sentence gains in Murray et al. (2007); (3) therapeutic activity at both sentence and discourse macrostructure levels of language, reporting word and sentence gains (Hickin et al. 2015). The four case studies scored moderately low for quality (27-43%) and similarly for completeness of reporting (33-50%). None of these studies received points for randomization, blinding of participants, blinding of assessors, treatment adherence, or replication, reflecting the fact that treatment of discourse is still in an early phase of investigation.

To some degree, it is unsurprising that most studies were incompletely reported. The TIDieR checklist was only published in 2014, and 19 studies in this review were published prior to this date. Pleasingly, most of the more important items were frequently reported, that is, treatment materials and procedures, and mode and dose of delivery. Less positively, theoretical rationales for the essential elements of the treatment were only partially or completely reported for two thirds of the studies. This makes it challenging to identify on how treatments are intended to work if authors have not reported on them. Overall, the least well reported information, according to the TIDieR, related to fidelity (only 6 studies reported this), tailoring (5 studies), and modification (4 studies). This gap in the evidence base is a cause for concern. Fidelity information would tell us the extent to which the intervention was delivered as planned and would aid in the evaluation of treatment

effects (Ledford & Gast, 2014). Tailoring and modification relate to changes to the treatment protocol for either an individual (tailoring) or for everyone (modification) and can include information about personalisation and individualised changes to dosage which are important for replication and implementation. Some of the studies with designs yielding lower quality evidence (i.e. those with no randomisation or control) were ranked highly in terms of treatment reporting. For example, the three scripting studies (two non-experimental case report series, and a group study with no control) were the highest scoring studies according to the TIDieR. The higher scores may be because participant data - which would provide information about mode of delivery as well as tailoring, modification and fidelity - are captured online through the technology involved in the intervention delivery. As none of these three studies met the study design criteria for quality rating (with the scales used in this review), we were unable to evaluate their design quality in this way.

Outcome measurement

The discourse treatment evidence base is characterized by a wide range of different ways to measure treatment outcome. A wealth of measures has been reported in the discourse measurement review literature, with recent reviews identifying more than 500 measures (Bryant et al., 2016; Pritchard et al., 2017). This is mirrored in the current review of discourse treatment in which we extracted 514 incidents of measure use from 25 studies. These outcome measures varied in terms of *what* was measured (words, sentences, discourse macrostructure, language ability more generally, functional communication and impact). The current review shows that the most commonly used measure across all discourse treatment research is a measure of single word production which has been extracted from a discourse sample. There are important implications from this for evaluating the efficacy of discourse treatment, given that a large body of data on the effect of discourse treatment seems to relate only to the *words* produced in discourse. Future research needs to ensure that we have measures in all discourse treatment studies that reflect the

linguistic complexity of discourse, including at minimum measures of word and sentence production and discourse macrostructure. This echoes the recommendation made by Dietz and Boyle (2018) that consensus be reached on a set of outcome measures for discourse treatment that address both microstructure and macrostructure.

The finding that a group of studies in this review used a similar measure (single words extracted from a discourse) potentially also raises a positive implication for future work. It might be argued that there is an opportunity here to strengthen future research by including a measure of words-in-discourse for inter-study comparative purposes. However, although this was a commonly used group of measures, the method by which words-in-discourse was assessed varied across studies. This comprised measures of CIUs, words, content words, narrative words, nouns, verbs, adjectives, target or topic words, accurate words, and open class words. These measures were also reported in multiple ways as counts, proportions, counts per minute, counts per utterance, and proportions per utterance or discourse. A final issue adding to the heterogeneity of this group of measures was the variety in discourse contexts used, that is, narrative, procedural, conversational, personal, and picture description. The choice of discourse context makes a difference in our evaluation of the effects of discourse treatment because we know that it makes a difference to the language produced by the speaker (Armstrong, 2000; Cruice et al., 2014).

Critique of the outcome measures reported raises concerns around measure *quality.* High quality measures are usually considered to be those which have been standardised; have some normative data (or within-aphasia norms); give objective quantitative data; and/or have robust psychometrics (Pritchard et al., 2017). In our review, these were the measures categorised as aphasia batteries, clinical assessments, and standardised word and sentence tasks. There were examples of studies from each treatment category which used a high-quality outcome measure, however only the word

production in discourse and multi-level (three-level) treatment studies used standardised measures of word and sentence production. Measures of lower quality, in this review, were the *bespoke* measures which were created specifically for individual studies (usually to compare scores on treated vs. untreated items). These are not widely available, may be difficult to replicate, and crucially may not result in reliable scores when measured across raters or time. These lower quality measures are prevalent in the discourse treatment evidence base.

Treatment activity

A key finding of this review was the lack of consensus about what constitutes discourse treatment. We found a wide range of therapeutic activities (see Table 6) targeting various levels of language (word, sentence discourse macrostructure), using various treatment approaches (client acts, such as production and practice, paper-based metalinguistic tasks communicative tasks technologically-supported tasks; single-level vs. multi-level treatment; individual vs. group treatment contexts), and varied treatment input (modelling, cueing, or feedback from the clinician). One likely reason for this variety is that this is an emerging field, in which it is appropriate to trial various treatment approaches and designs. An additional explanation for the variety is the dearth of theoretical rationales in the evidence base, which leads to a lack of cohesion in methodology. This lack may be due to an absence of available theoretical frameworks to underpin discourse treatment rationales, or lack of agreement about which theoretical framework to use. Because discourse treatments are complex, they require a theoretical rationale linked to components of intervention, in order to allow for a hypothesis about the expected mechanisms of change (Craig et al., 2008). This gap in the evidence base limits the potential for future work to build systematically on what we already know. Lack of detail about theoretical rationales and mechanisms of change is coupled with a shortfall in the completeness of treatment reporting in the discourse treatment evidence base, limiting the potential for replication and/or implementation of those treatments which have been shown to be

efficacious. Equally, clear description of an intervention facilitates understanding of its components and the mechanism for change (Yamato et al., 2016).

Treatment efficacy

Despite the limitations in the evidence base, there are clear positive signs in the review that discourse treatment is efficacious, with most of the reviewed studies (88%) reporting improvement in discourse production as a result of therapy. The most common benefit related to the production of words, with 21 studies reporting change at this level. Changes in sentence production and discourse macrostructure were Less frequently assessed and reported. Ten studies reported gains in sentence production (three reporting improved scores in bespoke or standardised tasks, and seven reporting improved sentence production in discourse). Only six of the 25 discourse treatment studies reported gains in discourse macrostructure (Carragher et al.; Dietz et al.; Osiejuk; Penn & Beecham; Whitworth; and Whitworth et al.). Five of these studies employed treatments explicitly targeting discourse macrostructure or multi-level therapies. However, there are many limitations in this efficacy evidence. Inconsistent patterns of assessment make it impossible to compare treatment approaches, and know whether treatment at one level may be effective in bringing about change at another. For example, enhancing word production in discourse may have knock-on effects for grammatical integrity and macrostructure, but in most studies this possibility remained unexplored. As has already been acknowledged, the strength of the evidence is also not strong. There were few group studies and many that lacked experimental control. Studies were inconsistent in their use of inferential statistics and maintenance of gains was not always assessed. There is the additional possibility of publication bias. Publication of case reports, which make up the bulk of the evidence here, may be more likely when results are positive. It is possible therefore that negative evidence is unavailable for scrutiny.

Implications for research and practice

Discourse treatments are an emerging area in aphasia rehabilitation, consisting of complex interventions. Our evidence synthesis reveals key areas for improvement that would strengthen the scientific evidence in this field. There is a need for more robustly planned research that employs high quality study designs, explores interventions that intentionally target all three linguistic levels of discourse, and which includes assessment of outcomes across all levels. Best practice with complex interventions (Craig et al. 2008) is to develop them systematically, using the best available evidence and theory. In terms of linguistic theory relating to discourse, there are diverse perspectives to draw from (see Armstrong, 2000) which creates a real challenge, as evidenced by the lack of reference to theory in the majority of papers in this review. Those few that explicitly mentioned a theoretical framework for discourse macrostructure (Carragher et al., 2015; Whitworth, 2010; Whitworth et al., 2015) drew on variety of sources. These included Halliday and Hasan (1976) for cohesion, and Labov (1972), Rumelhart (1975), and Stein and Glenn (1979) for story grammar. Additionally, then, we need more discourse treatment research which is underpinned by an explicit theoretical rationale, and the field could benefit from some consensus about which theoretical frameworks to use.

There are some positive findings in the review which have implications for clinical practice. Across the papers reviewed, single word production almost always improved regardless of treatment type delivered. This finding indicates that discourse gains might be achieved with relatively modest adaptations to existing treatments which should encourage clinicians to work with familiar word treatments, extending them to include treatment activity in sentence and discourse macrostructure contexts. In contrast to word production, sentence and discourse macrostructure required explicit treatment to achieve gains. This finding highlights the need for future research to develop new multi-level discourse interventions. Two of the review papers (Hoover et al., 2015; Whitworth et al., 2015) offer useful insights here and provide some guidance about how this might be done.

Conclusion

Discourse treatment is an emerging field in aphasia rehabilitation research with the vast majority of studies published in the last 13 years. The evidence base is limited by heterogeneity – in research design, focus and specification of treatment, and outcome measures – and synthesis is challenging to achieve. Despite this, there is a clear indication that discourse has become more prominent and is occupying a new position in clinical research. This is threefold: discourse as outcome measurement context; discourse as treatment context; and discourse macrostructure as treatment goal. Discourse is the context in which researchers measured treatment effects, suggesting increased recognition of more naturalistic contexts in outcome measurement. Word production extracted from discourse should now be considered a core outcome in intervention studies. Discourse is the context employed in all treatments reviewed in this paper, enlivening therapy with more picture description, storytelling, riddles, fables and games. Surely this enriched and often personalised learning context can only be beneficial for the clients to whom we provide services. Finally, discourse is now an intentional goal of treatment as macrostructure, with some early indication here that this warrants much more attention in future research. All such findings move us collectively closer to the goal of enabling people with aphasia to be able to talk more in their everyday lives.

References

Armstrong, E. (2000) Aphasic discourse analysis: The story so far. Aphasiology, 14(9), 875-892.

Antonucci, S. (2009). Use of semantic feature analysis in group aphasia treatment. *Aphasiology*, *23*, 854-866, DOI:10.1080/02687030802634405

Australian Aphasia Rehabilitation Best Practice Statement (2014) [online]. Available: http://www.aphasiapathway.com.au/?name=About-the-statements [Accessed 7th October 2019]

Baker, E. (2012). Optimal intervention intensity. *International Journal of Speech-Language Pathology*, *14*(5), 401-409.

Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane database of systematic reviews*, (6).

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics*, *30*(7), 489-518.

Canadian Stroke Best Practice Recommendations: Guidelines update (2015) [online]. Available: http://journals.sagepub.com/doi/pdf/10.1177/1747493016643553 [Accessed 7th October 2019]

Caplan, D., & Hanna, J. E. (1998). Sentence production by aphasic patients in a constrained task. *Brain and Language*, *63*(2), 184-218.

Carragher, M., Karen Sage, K., & Conroy, P. (2015). Preliminary analysis from a novel treatment targeting the exchange of new information within storytelling for people with nonfluent aphasia and their partners, *Aphasiology, 29* (11), 1383-1408, DOI: 10.1080/02687038.2014.988110

Carlomagno, S. & Losanno, N. (1991). Expressive language recovery or improved communicative skills: Effects of p.a.c.e. therapy on aphasics' referential communication and story retelling, *Aphasiology, 5*, 419-424, DOI: 10.1080/02687039108248544

Cherney, L., Halper, A., Holland, A. & Cole, R. (2008). Computerized Script Training for Aphasia: Preliminary Results. *American Journal of Speech-Language Pathology,17*, 19 – 34, DOI: 10.1044/1058-0360(2008/003)

Cherney, L. & Halper, A. (2008). Novel Technology for Treating Individuals with Aphasia and Concomitant Cognitive Deficits. *Topics in Stroke Rehabilitation, 15* (6), 542-554, DOI: 10.1310/tsr1506-542.

Cherney, L. (2010a). Oral Reading for Language in Aphasia (ORLA): Evaluating the efficacy of computer-delivered therapy in chronic nonfluent aphasia. *Topics in Stroke Rehabilitation, 17* (6), 423- 431, DOI: 10.1310/tsr1706-423.

Cherney, L. (2010b). Oral reading for language in aphasia: Impact of aphasia severity on cross-modal outcomes in chronic nonfluent aphasia. *Seminars in Speech and Language, 31* (1), 42-51, DOI: 10.1055/s-0029-1244952.

Cherney, L. R., Simmons-Mackie, N., Raymer, A., Armstrong, E., & Holland, A. (2013). Systematic review of communication partner training in aphasia: Methodological quality. International Journal of Speech-Language Pathology, 15(5), 535-545.

Cochrane Consumer Network – Levels of Evidence page (n.d.) retrieved from https://consumers.cochrane.org/levels-evidence 7th October 2019.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ, 337, a1655.

Cruice, M., Pritchard, M., & Dipper, L. (2014). Verb use in aphasic and non-aphasic personal discourse: What is normal?. *Journal of Neurolinguistics*, *28*, 31-47.

Davis, G. A., & Wilcox, M. J. (1981). Incorporating parameters of natural conversation in aphasia treatment: PACE therapy. In R. Chapey (Ed.), *Language intervention strategies in adult aphasia* (pp. 169–193). Baltimore, MD: Williams & Wilkins.

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: have we reached the tipping point?. *Aphasiology*, *32*(4), 459-464.

Dietz, A., Vannest, J., Maloney, T., Altaye, M., Holland, S. & Szaflarski, J. (2018). The feasibility of improving discourse in people with aphasia through AAC: clinical and functional MRI correlates. *Aphasiology*, *32* (6), 693-719, DOI: 10.1080/02687038.2018.1447641

Ellis, C., Henderson, A., Wright, H. H., & Rogalski, Y. (2016). Global coherence during discourse production in adults: A review of the literature. International journal of language & communication disorders, 51(4), 359-367.

Falconer, C. & Antonucci, S. (2012). Use of semantic feature analysis in group discourse treatment for aphasia: Extension and expansion. *Aphasiology, 26* (1), 64-82, DOI: 10.1080/02687038.2011.602390

Frattali C. M., Thompson C. K., Holland A. L., Wohl C. B., Ferketic M. M. (1995). American speech-language hearing association functional assessment of communication skills for adults. ASHA Fulfilment Operations, Rockville, MD.

Galletta, E. & Vogel-Eyny, A. (2015). Translational treatment of aphasia combining neuromodulation and behavioral intervention for lexical retrieval: Implications from a single case study. *Frontiers in Human Neuroscience, 9,* 1-13, DOI: 10.3389/fnhum.2015.00447

Goodglass, H. and Kaplan, E. (1972) The Assessment of Aphasia and Related Disorders. Henry Klimpton, London.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Boston diagnostic aphasia examination: Stimulus cards* (3rd ed.). London, Philadelphia. Lippincott Williams & Wilkins.

Gordon, J. (2007). A contextual approach to facilitating word retrieval in agrammatic aphasia. *Aphasiology, 21*, 643- 657, DOI: 10.1080/02687030701192141

Goral, M. & Kempler, D. (2009). Training verb production in communicative context: Evidence from a person with chronic non-fluent aphasia. *Aphasiology, 23* (12), 1383- 1397, DOI: 10.1080/02687030802235203

Greenwood, A., Grassly, J., Hickin, J. & Best, W. (2010). Phonological and orthographic cueing therapy: A case of generalised improvement. *Aphasiology, 24* (9), 991-1016, DOI: doi.org/10.1080/02687030903168220.

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. London: Longmans.

Helm-Estabrooks, N. (1981). Helm Elicited Language Program for Syntax Stimulation. Austin, Texas: Exceptional Resources Inc.

Helm-Estabrooks N. (2001). Cognitive Linguistic Quick Test (CLQT): Examiner's Manual. The Psychological Corporation, San Antonio, TX.

Hickin, J., Mehta, B., Dipper, L. (2015). To the sentence and beyond: a single case therapy report for mild aphasia, *Aphasiology, 29* (9), 1038-1061, DOI: 10.1080/02687038.2015.1010474

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., ... & Lamb, S. E. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, *348*, g1687

Holland, A. L. (1980). *Communicative abilities in daily living (CADL): A test of functional communication for aphasic adults*. Baltimore, MD: University Park Press.

Hoover, E., Caplan, D., Waters, G. & Budson, A. (2015). Effects of impairment-based individual and socially oriented group therapies on verb production in aphasia. *Aphasiology, 29* (7), 781-798, DOI: 10.1080/02687038.2014.989953

Hoover, E., Caplan, D., Waters, G. & Carney, A. (2017). Communication and quality of life outcomes from an interprofessional intensive, comprehensive, aphasia program (ICAP). *Topics in Stroke Rehabilitation, 24* (2), 82-90, DOI: 10.1080/10749357.2016.1207147.

Kagan, A., Simmons-Mackie, N., Victor, J. C., Carling-Rowland, A., Hoch, J., Huijbregts, M., et al. (2011). Assessment for Living with Aphasia. (ALA). Toronto, ON: Aphasia Institute.

Kaplan, E. F., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test*. Philadelphia, PA: Lea & Febiger.

Kearns, K. P. (1986). Systematic programming of verbal elaboration skills in chronic Broca's aphasia. In R. C. Marshall (Ed.), *Case studies in aphasia rehabilitation* (pp. 225–244). Austin, TX: Pro-Ed.

Kertesz, A. (1982). *The Western Aphasia Battery*, New York: Grune & Stratton, Inc.

Labov, W. (1972). The transformation of experience in narrative syntax. In W. Labov, *Language and the inner city.* Philadelphia: University of Pennsylvania.

LaPointe, L. L., & Horner, J. (1979). *Reading comprehension battery for aphasia*. Oregon: CC Publications.

Ledford, J. R., & Gast, D. L. (2014). Measuring procedural fidelity in behavioural research. *Neuropsychological Rehabilitation*, *24*(3-4), 332-348.

Lee, J., Kaye, R. & Cherney, L. (2009). Conversational script performance in adults with non-fluent aphasia: Treatment intensity and aphasia severity. *Aphasiology, 23*(7), 885-897, DOI: 10.1080/02687030802669534.

Linnik, A., Bastiaanse, R., & Höhle, B. (2016). Discourse production in aphasia: A current review of theoretical and methodological challenges. *Aphasiology*, 30(7), 765-800.

Lomas J, Pickard L, Bester S, Elbard H, Finlayson A, Zoghaib C. (1989) The communicative effectiveness index: development and psychometric evaluation of a functional measure of adult aphasia. *Journal of Speech and Hearing Research,* 54, 113–124

Maher, C.G., Sherrington, C., Herbert, R.D., Moseley, A.M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. Physical Therapy, 8, 713–721.

Marini, A., Caltagirone, C., Pasqualetti, P., Carlomagno, S. (2007). Patterns of language improvement in adults with non-chronic non-fluent aphasia after specific therapies. *Aphasiology, 21* (2), 164- 186, DOI: 10.1080/02687030600633799

McCall, D., Virata, T., Linebarger, M.C., Berndt, R.S. (2009). Integrating technology and targeted treatment to improve narrative production in aphasia: A case study. *Aphasiology, 23* (4), 438- 461, DOI: 0.1080/02687030701818000

Murray, L., Timberlake, A. & Eberle, R. (2007). Treatment of underlying forms in a discourse context. *Aphasiology, 21* (2), 139- 163, DOI: 10.1080/02687030601026530

Milman, L., Clendenen, D. & Vega-Mendoza, M. (2014). Production and integrated training of adjectives in three individuals with nonfluent aphasia. *Aphasiology, 28* (10), 1198- 122, DOI: 10.1080/02687038.2014.910590

NICE (2018). Glossary [online]. Accessed: https://www.nice.org.uk/glossary?letter=r [25th October 2018]

Nickels, L., McDonald, B. & Mason, C. (2016). The impact of group therapy on word retrieval in people with chronic aphasia. *NeuroRehabilitation, 39* (1), 81-95, DOI: 10.3233/NRE-161340.

Osiejuk, E. (1991). Short report discourse exercises in aphasia therapy: A case study, *Aphasiology, 5*, 443- 446, DOI: 10.1080/02687039108248549

OCEBM (n.d.) Levels of Evidence Working Group. "The Oxford Levels of Evidence 2". Oxford Centre for Evidence-Based Medicine. Retrieved from https://www.cebm.net/index.aspx?o=5653 7th October 2019

Paul, D. R., Frattali, C. M., Holland, A. L., Thompson, C. K., Caperton, C. J. and Slater, S. C. (2004). *Quality of Communication Life Scale*, The American Speech-Language-Hearing Association. Rockville, MD.

Perdices, M. (2015). The risk of bias in N-of-1 trials (RoBiNT) Scale: an expanded manual for the critical appraisal of single-case reports. Sydney, Australia: Author.

Pollock, A., St George, B., Fenton, M., & Firkins, L. (2012). Top ten research priorities relating to life after stroke. *The Lancet Neurology*, *11*(3), 209.

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2017). Reviewing the quality of discourse information measures in aphasia. International Journal of Language & Communication Disorders, 52(6), 689-732.

PsychBITE (2018). PEDro-P Scale, Rating Scale for Randomised and Non-Randomised Controlled Trials [online]. Available: http://www.psycbite.com/web/cms/content/docs/The_PEDro-P_Scale.pdf [accessed 25th October 2018]

Roach A, Schwartz MF, Martin N, Grewal RS, Brecher A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology*. 24:121–133.

Rochon, E., Laird, L., Bose, A. & Scofield, J. (2005). Mapping therapy for sentence production impairments in nonfluent aphasia. *Neuropsychological Rehabilitation, 15* (1), 1-36, DOI: 10.1080/09602010343000327.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow & A. Collins (Eds.), Representation and understanding: Studies in cognitive science (pp. 211–236). New York, NY: Academic Press.

Stein, N., & Glenn, C. (1979). An analysis of story comprehension in elementary school children. Hillsdale, NJ: Erlbaum.

Swinburn, K., Porter, G. and Howard, D. (2005). *The Comprehensive Aphasia Test*, Hove, UK: Psychology Press.

Swinburn, K. and Byng, S. 2006. *The Communication Disability Profile*, London: Connect Press

Tate, R., Rosenkoetter, U., Wakim, D., Sigmundsdottire, L., Doubleday, J., Togher, L., McDonald, S., & Wambaugh, J.L., Nessler, C., Wright, S. (2013). Modified response elaboration training: Application to procedural discourse and personal recounts. *American Journal of Speech-Language Pathology, 22* (2), S409- S4425, DOI: 10.1044/1058-0360(2013/12-0063).

Ulatowska, H. K., Freedman-Stern, R., Doyel, A. W., Macaluso-Haynes, S., & North, A. J. (1983). Production of narrative discourse in aphasia. *Brain and language*, *19*(2), 317-334.

Walker, G. M., & Schwartz, M. F. (2012). Short-form Philadelphia naming test: Rationale and empirical evaluation. American Journal of Speech-Language Pathology, 21, S140–153. doi:10.1044/1058-0360(2012/11-0089).

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

Webster, J., Whitworth, A., & Morris, J. (2015). Is it time to stop "fishing"? A review of generalisation following aphasia intervention. *Aphasiology*, *29*(11), 1240-1264. DOI: 10.1080/02687038.2015.1027169

Wechsler, D. (2008). Wechsler adult intelligence scale–Fourth Edition (WAIS–IV). *San Antonio, TX: NCS Pearson*, *22*, 498.

Whitworth, A. (2010). Using narrative as a bridge: Linking language processing models with real-life communication. *Seminars in Speech and Language, 31* (1), 64-75, DOI: 10.1055/s-0029-1244954

Whitworth, A., Leitão, S., Cartwright, J., Webster, J., Hankey, G.J., Zach, J., Howard, D., Wolz, V. (2015). NARNIA: a new twist to an old tale. A pilot RCT to evaluate a multilevel approach to improving discourse in aphasia. Aphasiology 29 (11), 1345-1382, 10.1080/02687038.2015.1081143.

Webster, J., Whitworth, A., & Morris, J. (2015). Is it time to stop "fishing"? A review of generalisation following aphasia intervention. *Aphasiology*, *29*(11), 1240-1264.

Yamato, T., Maher, C., Saragiotto, B., Hoffmann, T., & Moseley, A. (2016). How completely are physiotherapy interventions described in reports of randomised trials? Physiotherapy Research International, 102, 121–126.

Table 1

Inclusion and exclusion criteria, and list of eligible papers

Criteria	Definition
Aphasia	LH stroke, damage affecting language spheres of brain. Acquired aphasia- describing people aged 18+. Exclusion: U18, progressive or TBI, dementia, other pathology.
Primary data reported in a peer-reviewed journal	Primary data about study participants reported in a peer-reviewed journal. Exclusion: reviews, conference papers, theoretical papers, unpublished manuscripts; book chapters; unpublished PhD or Master's level dissertations or theses.
English language	Inclusion: Abstract and article available in English language Exclusion: Abstract or full text published in a language that is not English
Discourse	Connected speech beyond a single sentence, for a specific purpose (e.g., describing, storytelling, arguing, persuading, advising). Monologues and dialogues, and conversation.
Direct SLT intervention	Clinical intervention, working directly with PWA. Exclusion= commentaries, reviews; work with significant others.
Discourse assessment	Assessment of one or more aspects of discourse, using established or less widely used methods (e.g., time, fluency, information, language, structure, using established or information elicitation)

Discourse targeted intervention	Targeted clinical intervention, where a component involves working with discourse within the intervention ('connected speech beyond a single simple sentence') with the aim of improving it.

List of papers meeting the criteria and included in the review:

1. Antonucci (2009).
2. Carlomagno & Losanno (1991).
3. Carragher (2015).
4. Cherney (2010).
5. Cherney & Halper (2008).
6. Cherney, Halper, Holland & Cole (2008).
7. Cherney (2010)
8. Dietz, Vannest, Maloney, Altaye, Holland & Szaflarski (2018).
9. Falconer & Antonucci (2012).
10. Goral & Kempler (2009).
11. Gordon (2007).
12. Greenwood, Grassly, Hickin & Best (2010).
13. Hickin, Mehta & Dipper (2015).
14. Hoover, Caplan, Waters & Budson (2015).
15. Lee, Kaye & Cherney (2009).
16. Marini, Caltagirone, Pasqualetti & Carlomagno (2007)
17. McCall, Virata, Linebarger & Berndt (2009).
18. Milman, Clendenen & Vega-Mendoza (2014).
19. Murray, Timberlake & Eberle (2007).
20. Nickels, McDonald & Mason (2016).
21. Osiejuk (1991)
22. Penn & Beecham (1992).
23. Wambaugh, Nessler & Wight (2013).
24. Whitworth (2010).
25. Whitworth, Leitão, Cartwright, Webster, Hankey, Zach, Howard & Wolz (2015).

Table 2

Discourse Treatment Papers grouped by Treatment Type

Treatment Group	Definition	Studies
Word production in discourse (n=5)	Therapeutic activities focus on improving speakers' production of single words	Antonucci, 2009; Falconer & Antonucci, 2012; Greenwood et al., 2010; Gordon, 2007; Nickels et al., 2016
Sentence production in discourse (n=5)	Therapeutic activities focus on improving speakers' production of sentences or utterances	Cherney, 2010a; Cherney, 2010b; Goral & Kempler, 2009; Murray et al., 2007; Wambaugh et al., 2013
Discourse macrostructure (n=2)	Therapeutic activities focus on improving the overall structure and information a speaker communicates in discourse	Osiejuk, 1991; Carragher et al. 2015
Discourse scripts (n=3)	Therapeutic activities focus on improving production of a whole script	Cherney et al. 2008; Cherney & Halper, 2008; Lee et al., 2009
Multi-level (2 levels) (n=5)	Therapeutic activities focus on any combination of two of the levels: single word, sentence, discourse	Carlomagno & Lassanno, 1991; Hickin et al., 2015; Marini et al., 2007; Milman et al., 2014; Penn & Beecham, 1992

(3 levels) (n=3)	Therapeutic activities focus on all three levels: single word, sentence, discourse	Hoover et al., 2015; Whitworth, 2010; Whitworth et al., 2015
No consensus[5] (n=2)	Consensus about the focus of therapeutic activity(ies) could not be reached	Dietz, Vannest, Maloney, Altaye, Holland & Szaflarski, 2018; McCall, Virata, Linebarger & Berndt, 2009

[5] Consensus could not be reached for categorising these two studies into one of the treatment groups. Although they describe technological support for language at multiple levels (word, sentence, discourse macrostructure), we could not reach agreement about whether there was therapeutic activity at each level. The decision was therefore made to leave them uncategorised.

Table 3

Methodological Quality Ratings for Clinical Trials and Single-Case Experimental Designs.

Clinical Trials			
Author (year)			**PEDro-P Score (out of 10)**
Cherney (2010)			6 (60%)
Dietz, Vannest, Maloney, Altaye, Holland, & Szaflarski (2018)			6 (60%)
Whitworth, Leitão, Cartwright, Webster, Hankey, Zach, Howard, & Wolz (2015)			7 (70%)
Single-Case Experimental Designs			
	RoBiNT Score		
Author (year)	**Internal Validity Subscale (out of 14)**	**External Validity Subscale (out of 16)**	**Total Score (out of 30)**
Goral & Kempler (2009)	3	10	13 (43%)
Greenwood, Grassly, Hickin, & Best (2010)	0	8	8 (27%)
Hickin, Mehta, & Dipper (2015)	0	11	11 (37%)

Murray, Timberlake, & Eberle (2007)	3	10	13 (43%)

Table 4
TIDieR overview of aphasia discourse treatment studies
For details of the Items, please see Appendix 2
KEY: + = reported in the article; / = partial reporting in the article

Article	**Item 1**	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Item 6**	**Item 7**	**Item 8**	**Item 9**	**Item 10**	**Item 11**	**Item 12**	**Total +**
Word Production in Discourse Treatment Studies (n=5)													
Antonucci (2009)	+	+/	+	+		+		+	/				6
Falconer & Antonucci (2012)	+	+/	+	+	/	+		+					6
Gordon (2007)	+		+	+		+		+		+			6
Greenwood et al. (2010)	/		+	+		+	/	+					4
Nickels et al. (2016)	/	+/	+	+	+	+		+		/			6
Sub-total	*3*	*3*	*5*	*5*	*1*	*5*	*0*	*5*	*0*	*1*	*0*	*0*	*N/A*
Sentence Production in Discourse Treatment Studies (n=5)													
Cherney (2010a)	+	+	/	+	/	+		+		/			5
Cherney (2010b)	+	+	/	+	/	+	/	+		?			5

Article	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Total +
Goral & Kempler (2009)	+	+	+	+		+		+					6
Murray et al. (2007)	+	+	+	+		+		+					6
Wambaugh et al. (2013)	+		+	+	+	+	+	+	+				8
Sub-total	*5*	*4*	*3*	*5*	*1*	*5*	*1*	*5*	*1*	*0*	*0*	*0*	*N/A*
Discourse Macrostructure Treatment Studies (n=2)													
Carragher et al. (2015)	+	+	+	+	+	+		+	+				8
Osiejuk (1991)			+	+	/	/	+	+					4
Sub-total	*1*	*1*	*2*	*2*	*1*	*1*	*1*	*2*	*1*	*0*	*0*	*0*	*N/A*
Discourse Scripts Treatment Studies (n=3)													
Cherney et al. (2008)	+	+	+	+	+	+	+	+	+	+	+	+	12
Cherney & Halper (2008)	+	+	+	+	+	+	+	+	+		+	+	11
Lee et al. (2009)	+	+	+	+	+	+	+	+	+	+	+	+	12
Sub-total	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*2*	*3*	*3*	*N/A*

Article	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Total +
Multi-level Treatment Studies – 2 linguistic levels (n=5)													
Carlomagno et al. (1991)	+	+	+	+		+		+					6
Hickin et al. (2015)			+	+		+	+	+					5
Marini et al. (2007)	+	+	+	+	+	+	+	+			+		9
Milman et al. (2014)	+	+	+	+		+		+		+		+	10
Penn & Beecham (1992)			+	+	+	+		+					5
Sub-total	*3*	*3*	*5*	*5*	*2*	*5*	*2*	*5*	*0*	*1*	*1*	*1*	*N/A*
Multilevel Treatment Studies – 3 linguistic levels (n=3)													
Hoover et al. (2015)	/	+/	/	/	+	+		+			+		5
Whitworth et al. (2015)	+	+	+	+	+			+			+		7
Whitworth (2010)			+	+		/		+					3
Sub-total	*1*	*2*	*2*	*2*	*2*	*0*	*0*	*3*	*0*	*0*	*2*	*0*	*N/A*
No consensus (n=2)													

Article	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Total +
Dietz et al. (2018)		+	+	+		+	+	+			/	+	7
McCall et al. (2009)	+		+	+		+	+	+					6
Sub-total	*1*	*1*	*2*	*2*	*0*	*2*	*2*	*2*	*0*	*0*	*0*	*1*	*N/A*
GRAND TOTAL	***17***	***17***	***22***	***24***	***10***	***22***	***9***	***25***	***5***	***4***	***6***	***5***	***-***

Table 5

Incidents of Outcome Measures Use for each category of Treatment

Treatment Group (number of studies)	**Number of incidents of measure use (*mean per study*)**
Word production in discourse (n = 5)	82 *(16)*
Sentence production in discourse (n = 5)	116 *(23)*
Discourse macrostructure (n = 2)	44 *(22)*
Discourse scripts (n = 3)	22 (*7*)
Multi-level (2 linguistic levels) (n = 5)	163 *(33)*
Multi-level (3 linguistic levels) (n = 3)	56 *(19)*
No consensus (n = 2)	31 *(16)*
TOTAL	514

Table 6

Description of Treatment for each Paper

Article	***Description of treatment***
Single word production in discourse treatment studies	
Antonucci (2009)	Treated nouns using semantic feature analysis (SFA); group context, involving picture description (of increasing complexity) to other members using PACE-like activities for the majority of treatment sessions. The study used *Descripto-Cards® — Advanced: For Functional Language and Cognition*, *Narrative Story Cards* (Helm-Estabrooks & Nicholas, 2003), and picture stimuli from the *Melodic Intonation Therapy Program* (Helm-Estabrooks, Nicholas, & Morgan, 1989).
Falconer & Antonucci (2012)	First author treated nouns using SFA; group context, involving picture description (of increasing complexity) to other members using PACE-like activities, for the majority of treatment sessions. *Descripto-Cards® — Advanced: For Functional Language and Cognition*, *Narrative Story Cards* (Helm-Estabrooks & Nicholas, 2003), and picture stimuli from the *Melodic Intonation Therapy Program* (Helm-Estabrooks, Nicholas, & Morgan, 1989).
Gordon (2007)	Treated nouns using 'semantic treatment' inspired by SFA and a contextual treatment delivered as listening to a story then retelling the story with clinician cueing, in one treatment condition. Bespoke stimuli and face-to-face treatment.
Greenwood et al. (2010)	Treated face-to-face with 40 personally chosen nouns (amongst other naming stimuli: 200 black and white naming cards) using SFA and phonological and orthographic cueing therapy, delivered in phases, with phase 2 incorporating conversation.
Nickels et al. (2016)	Treatment delivered by second author. Target production in response to focused discussion in group treatment, combined with Repetition In the Presence of a Picture (RIPP) in a home programme. The group aphasia treatment was based on discussion/ response to questions on a different observed 30-minute documentary each session; Bespoke stimuli.
Sentence production in discourse treatment studies	
Cherney (2010a)	ORLA (Oral reading for language in aphasia); treatment (face-to-face) involved repetitive multimodal stimulus targets written on index cards or on a computer screen. Participants saw a written stimulus at the same time as hearing it, and then read it aloud with the therapist, with stimulus complexity tailored to client level and ranging from short sentences of three words to brief paragraphs of 30 words.
Cherney (2010b)	ORLA (Oral reading for language in aphasia); treatment (face-to-face or computer delivered) involved repetitive multimodal stimulus targets written on index cards or on a computer screen. Participants saw a written stimulus at the same time as hearing it, and then read it aloud with the therapist, with stimulus complexity tailored to client level and ranging from short sentences of three words to brief paragraphs of 30 words.
Goral & Kempler (2009)	CIAT (constraint induced aphasia therapy); face-to-face; verbs selected by relevance to participants' life and communication needs, relating to construction, home, recreation and semantic lightness. The CIAT focused on verbal communication, with the participant repeating and reading sentences with SVO structures, and then completing discourse activities using the SVO structures: describing pictures, playing 'go fish', describing picture cards, making phone calls using a script, generating stories, recounting episodes, and having conversations.

Murray et al. (2007)	Treatment of Underlying forms (TUF); treated face-to-face; spoken and written sentence production; clinician modelling target sentence; using word and phrase cards with active and passive sentences; client copying the sentence in written form, and identifying thematic roles - verb, agent, theme; client rearranging an anagram of the sentence; and the client supplying the sentence in written format, followed by a discourse training module, where the client used the target sentence structure within a five-sentence factual or imaginative story about a current newspaper or magazine article.
Wambaugh et al. (2013)	Modified Response Elaboration Training (RET); treated face-to-face; clients' self-generated productions in response to a picture, which were then elaborated upon by the clinician. Clinician requested personal and procedural discourses ('tell me in detail how you would go about…' and 'tell me something about anything that you would like to talk about'), with the client's original utterances then reinforced, modelled, and additional prompts and forward chains provided. RET was tailored to client through use of personal discourse utterances and personally relevant procedural utterances as the basis for the intervention.
Discourse macrostructure studies	
Carragher et al. (2015)	Therapy was delivered face-to-face by the first author, a speech and language therapist. Target: video clips from YouTube involving minimal or no use of language, achieved through a set of treatment activities targeting the exchange of new information within storytelling. Participants reviewed a baseline video of themselves storytelling, and reflected on behaviours, strategies, and set personalized goals, with PWA goals focused on story grammar and CP goals focused on strategies to repair breakdowns. Sessions with people with aphasia focused on improving their thinking for speaking, segmenting narrative into events (broadly beginning, middle, end); identifying the main referent; creating a visual record, and modelling of syntactic constructions (agent verb). The cueing and prompts were maximal at telling one, and reduced over tellings 2 and 3 with the Conversation Partner. Finally, the dyad watched the YouTube video together, reflected, and identified strategies to practise at home, within daily conversation.
Osiejuk (1991)	The targets included narrative and procedural discourses, some chosen by the client, and simple life events, through discourse production exercises. A range of stimuli were used, including pictures (e.g., single scenes and simple and complex stories), and verbal (e.g., short fables, simple life events). Following this, the client composed a description, providing information about the setting, an action with a sequence of events, a resolution, or steps within a procedure. The client then added detail to the discourses; retold the story from memory without visual or verbal cues; gave the general semantic theme (topic or title), gave examples of a similar situations (e.g. making scrambled eggs = a kind of dish preparation), and give interpretation, commentary, morals of stories and resolutions. No details were given about who the therapist was and it was implied that delivery of therapy was face-to-face.
Discourse script studies	
Cherney et al. (2008)	The targets in all three studies were personalised scripts (n=3 for each speaker, monologues or dialogues up to 20 turns long). Topics were chosen by the client, written by the SLP and agreed and modified collaboratively. Scripts were monologic or dialogic discourses, such as the client telling the story of his or her stroke or giving a short lecture. Therapy was delivered using AphasiaScripts™, a software program with virtual therapist. The scripts are inputted by a SLP, but the software is used by clients independently at home, choosing their practice
Cherney & Halper (2008)	
Lee et al. (2009)	

	components (single words, sentences, the entire script), and regularity of practice. The virtual therapist is programmed to produce natural speech with correct movements of the speech articulators, and clients' scripts are inputted into the software in written and audio formats, which are synced. This means therapy stimuli are delivered via audio (a recorded version of the script), visual (watch the therapist's oral motor movements; and seeing the written sentences on the screen, synced with the audio. Clients listen silently to the whole script with each word highlighted; practise component words and whole sentences repeatedly; and practise the whole script.
Multi-level studies (2 linguistic levels)	
Carlomagno et al. (1991)	Discourse level work in combination with single words (objects) using PACE/modified PACE and picture description/picture sequences. Targets progressed from communicating single objects and scenes; riddle play using famous people; picture description; to more complex pictures and telling of stories in cartoon-like fashion, beginning with explicit requests of themes and alternative communication strategies, and fading these to vague requests for more detailed information.
Hickin et al. (2015)	Discourse level work in combination with sentence level targets using Story Grammar and picture sequences. Therapy began with combining sentences using a target connective; then combining a series of sentences as a narrative; showing picture sequences as a narrative description; and then producing personalized spontaneous narratives about a topic of interest such as the current storyline in the participant's favorite soap opera or the latest film she had seen. Therapy included explanation of sentence types and picture materials used in Teaching English as a Foreign Language, provided via weblink, and SentenceShaper books D and E.
Marini et al. (2007)	Discourse level work in combination with sentence level targets, delivered by two experienced speech therapists of the IRCCS Santa Lucia. They used the HELPSS method on sentences, and then used PACE with picture-description, stimulus response and story-telling/story completion activities. The speaker was encouraged to communicate more information in these activities using alternative communication strategies modelled by the clinician, and using explicit requests when appropriate.
Milman et al. (2014)	Discourse level work in combination with single words (adjectives) using modified Response Elaboration Training. Focused on photos depicting single word adjective production, also practised in sentences provided in an online appendix. Although the adjectives in this study were also practised in discourse, this study was coded therapeutic at single word and sentence levels only, because explicit shaping and feedback was provided at these levels. The group session gave participants the opportunity to practise the target structures.
Penn & Beecham (1992)	Discourse level work in combination with single words (single word naming) with self-monitoring. Focused on conversational strategies and personal discourse e.g., 'what did you do this morning?', identified by the client and the therapist; reflecting on video and audio recordings, and practised using culturally neutral picture description and story sequence tasks. Therapy delivered by 'a monolingual clinician/therapist'.
Multi-level studies (3 linguistic levels)	
Hoover et al. (2015)	The single word targets were verbs using an individual impairment-orientated treatment which was a modified version of VNeST. This was followed by a socially oriented group treatment using a conversational format, focusing on discourse (Elman, 2007; Elman & Bernstein-Ellis, 1999b; Simmons-Mackie &

	Chapey). Participants also engaged in language games, functional scripts and discussion using the verbs surrounding the nine conversational topics. Stimuli in this study were drawn from Toomey & Newman's (1992) 'Sequence Plus' sentence cue card and mindmaps and discourse generated using the topic headings detailed the verb lists (provided in the appendix). Face-to-face delivery by licensed SLPs or graduate students in speech-language pathology directly supervised by an SLP, and group treatments were provided by a licensed SLP.
Whitworth (2010)	Largely the same protocol as the RCT below. The single word targets were verbs, with focus on identifying main events, accessing verb argument structure, and story grammar, using mind maps. Face-to-face delivery by two qualified speech pathologists
Whitworth et al. (2015)	The single word targets were verbs, with focus on identifying main events, accessing verb argument structure, and story grammar, using mind maps. Participants progressed through picture sequences of up to five events, followed by different discourse genres involving situations close to everyday life, such as event recall, event planning, providing opinions. Intervention focused on identifying the main event/s in each picture, accessing the verb and the relevant nouns for each event, followed by creating a complete argument structure around each. The framework for narrative discourse involved sentences being organized around setting the scene (the beginning), the events taking place (the middle), and concluding the story (the end). Integration of word, sentence, and narrative levels was a key feature. This study encouraged targets tailored to each individual as lexical selection was appropriate to each participant's needs and ability; and discourses targeted included event recall, event planning, providing opinions, which are likely to have been personally relevant. Face-to-face delivery.
No consensus studies	
Dietz et al. (2018).	Targets were stories participants would feel comfortable with sharing during treatment, and with unfamiliar people. The study used a Dynavox VMAX AAC device with the personal narratives, selected from a choice of 6-8 personally relevant stories, programmed into the device. When possible, the researchers encouraged the participants to have a caregiver help to write out key ideas for each story. Participants told their stories, self-cuing word retrieval using various elements of the interface. Treatment was self-administered, following lab training, with weekly face to face check ins in both studies. Treatment was delivered in an individual context.
McCall et al. (2009)	Targets were multiclause sentences programmed into SentenceShaper software. There were two phases of treatment, with the first focusing on complex sentence production, and the second focusing on combining these sentences into a narrative, building up picture sequences, and then wordless picture books. Treatment was self-administered, following lab training, with weekly face to face check ins in both studies. Treatment was delivered in an individual context.

Table 7

Overview of Results for the Word Production in Discourse Studies (N=12 participants)

Measure	Antonucci, 2009 N=2			Falconer & Antonucci, 2012 N=4			Gordon, 2007 N=2			Greenwood et al., 2010 N=1			Nickels et al., 2016 N=3		
	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats
Word production extracted from discourse	✓	✓	Effect Sizes	✓	✓3 PWA	Effect Sizes	X	N/A	None	✓	✓	Descript.	X	N/A	Descript
Word errors from discourse	✓	✓	Descript.	✓	✓	Descript.				✓	✓	Descript.			
Standardized word production task (BNT)	X	N/A	Descript.	✓1 PWA	N/A	Descript.									
Bespoke word production task							✓	✓	Descript.	✓	✓	Non-parametric	X	N/A	Parametric
Aphasia battery (WAB)	✓ (>5 points)	N/A	Descript.	✓ (>5 points)	N/A	Descript.	✓*	✓	Descript.						
Other (impact)										✓ (CDP)	✓	Descript.			

*= substantial change

Table 8

Overview of Results for the Sentence Production in Discourse Studies (N=30 participants)

Measure	**Cherney 2010a & 2010b, N=25**			**Goral and Kempler, 2009 N=1**			**Murray, et al., 2007 N=1**			**Wambaugh, et al., 2013 N=3**		
	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats
Word production extracted from discourse	✓	N/A	Parametric	✓ Word and morphology	✓	Effect sizes	✓	✓	Descript.	✓ Procedural discourse X Personal discourse	✓ N/A	Effect sizes
Sentence production extracted from discourse							X	N/A	Descript.			
Bespoke sentence production task							✓	✓	Descript.			
Aphasia Battery (ies)	✓ (WAB)	N/A	Effect sizes	X (BDAE)	N/A	Descript.						
Other (social communication)				✓	N/A	Parametric						

Table 9

Overview of Results for the Discourse Macrostructure Studies (N=5 participants)

Measure	**Carragher et al., 2015 N=4**			**Osiejuk, 1991 N=1**		
	Change	Main.	Stats	Change	Main.	Stats
Word production extracted from discourse	✓	N/A	Descript.	✓	N/A	Descript.
Sentence production extracted from discourse				✓	N/A	Descript.
Discourse macrostructure	✓	N/A	Descript.	✓	N/A	Descript.
Aphasia Battery (BDAE)				✓	N/A	Descript.
Other (WAIS)				X	N/A	Descript.

Table 10

Overview of Results for the Discourse Script Studies (N=23 participants)

Measure	**Cherney & Halper, 2008 N=3**			**Cherney et al., 2008 N=3**			**Lee et al., 2009 N=17**		
	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats
Word production extracted from discourse	✓2 PWA (2/3 scripts)	N/A	Descript.	✓	N/A	Descript.	✓	N/A	Descript.
Aphasia Battery (WAB)	✓1 PWA	1PWA	Descript.	✓2 PWA	N/A	Descript.			
Other				✓1 PWA (QCL) **(CADL)**X	N/A	Descript.			

Table 11

Overview of Results for the Multi-level Studies with two levels of language treated (N=16 participants)

Measure	**Carlomagno et al., 1991 N=8**			**Hickin et al., 2015 N=1**			**Marini et al., 2007 N=3**			**Milman et al., 2014 N=3**			**Penn and Beecham, 1992 N=1**		
	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats
Bespoke word production task										✓2 PWA	✓	Effect sizes			
Standardized word production task (BNT)										✓2 PWA					
Bespoke sentence production task										✓2PWA	✓	Effect sizes			
Word production extracted from discourse	X (Cookie Theft) ✓ (Story retell)	N/A	Non-parametric	✓	N/A	Descript.	✓	✓ No stats	Parametric	✓2 PWA	N/A	Descript.			
Sentence production extracted from discourse				✓	N/A	Non-parametric	X	N/A	Parametric						
Discourse macrostructure				X	N/A	Descript.							✓	N/A	Descript.
Aphasia Battery (ies)				✓ (CAT picture)	N/A	Descript.	✓1 PWA (AAT) (	N/A	Descript.	✓2 PWA (WAB)	N/A	Descript.			
Other				✓ (QoL)	N/A	Descript.	✓1 PWA (CADL) ✓ (All) (FCP)	N/A	None	X (Raven's) ✓1 PWA (SSCAN)	N/A	Descript.	✓ (strat-egies)	N/A	Descript.

Table 12

Overview of Results for the Multi-level Studies with three levels of language treated (N=28)

Measure	**Hoover et al., 2015 N=12**			**Whitworth, 2010 N=2**			**Whitworth et al., 2015 N=14**		
	Change	Main.	Stats	Change	Main.	Stats	Change	Main.	Stats
Word production extracted from discourse	X	N/A	Parametric	✓1PWA	✓	Non-parametric	✓[?]	✓	Parametric
Standardised word production task	✓	✓	Parametric				✓[?]	N/A	
Sentence production extracted from discourse	✓	✓	Parametric	✓	✓	Non-parametric	✓[?]	✓	Parametric
Standardised sentence production task	✓	X	Parametric				X	N/A	
Discourse macrostructure				✓	✓	Non-parametric	✓[?]	X	Parametric
Other	✓ (ALA) ✓ (ASHA FACS)	N/A	Parametric						

[?] = within-group pre/post change, but no between-group difference

Table 13

Overview of Results for the No Consensus studies (N=13 participants)

Measure	**Dietz et al., 2018 N=12**			**McCall et al., 2009 N=1**		
	Change	Main.	Stats	Change	Main.	Stats
Word and morphology production extracted from discourse	✓	N/A	Parametric	✓	N/A	Non-parametric
Sentence production extracted from discourse	✓	N/A	Parametric	✓	N/A	Non-parametric
Discourse macrostructure	✓	N/A	Parametric			
Aphasia Battery (WAB)	✓	N/A	Parametric			

Figure 1

PRISMA Flow Diagram showing systematic search process

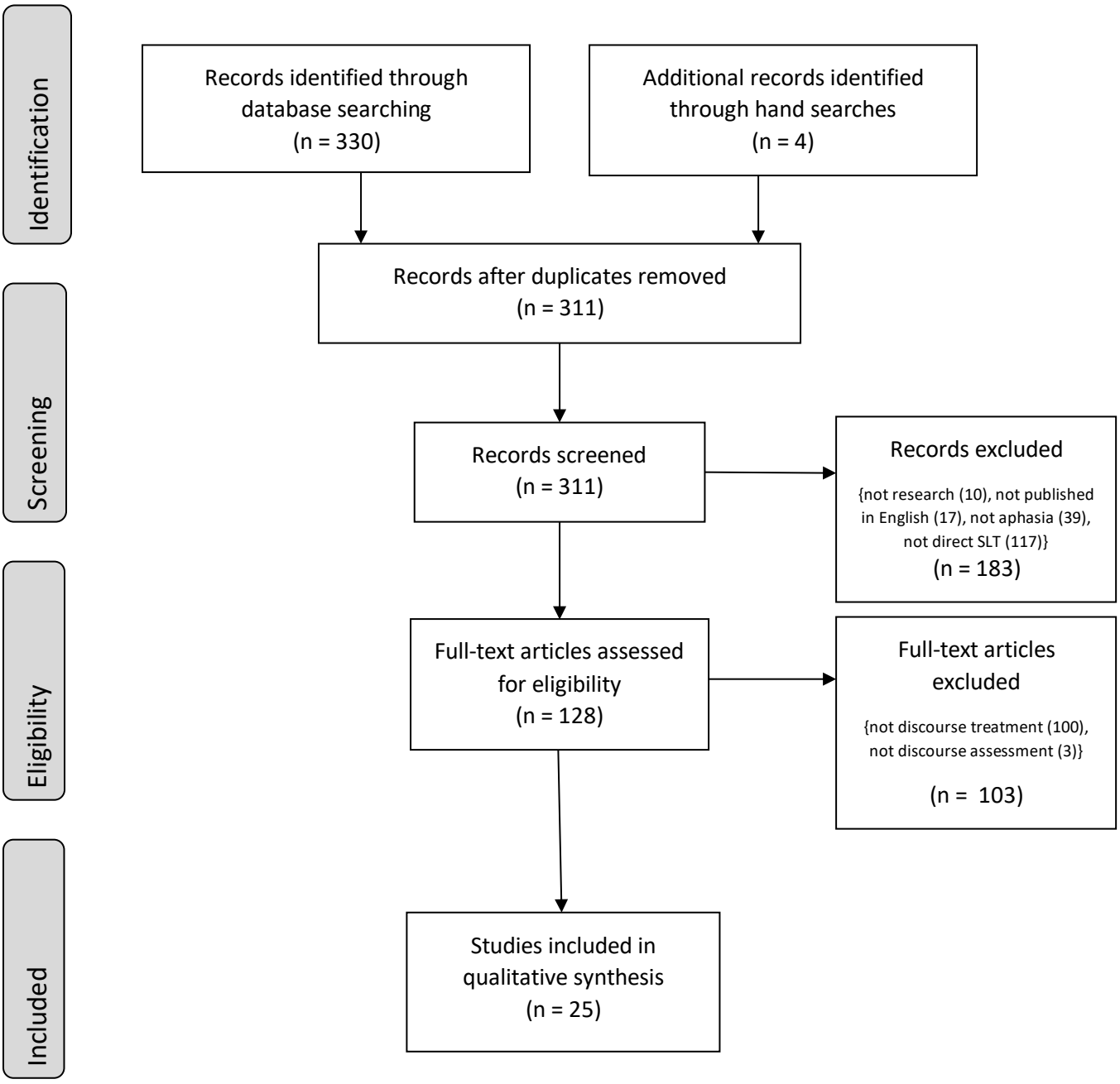
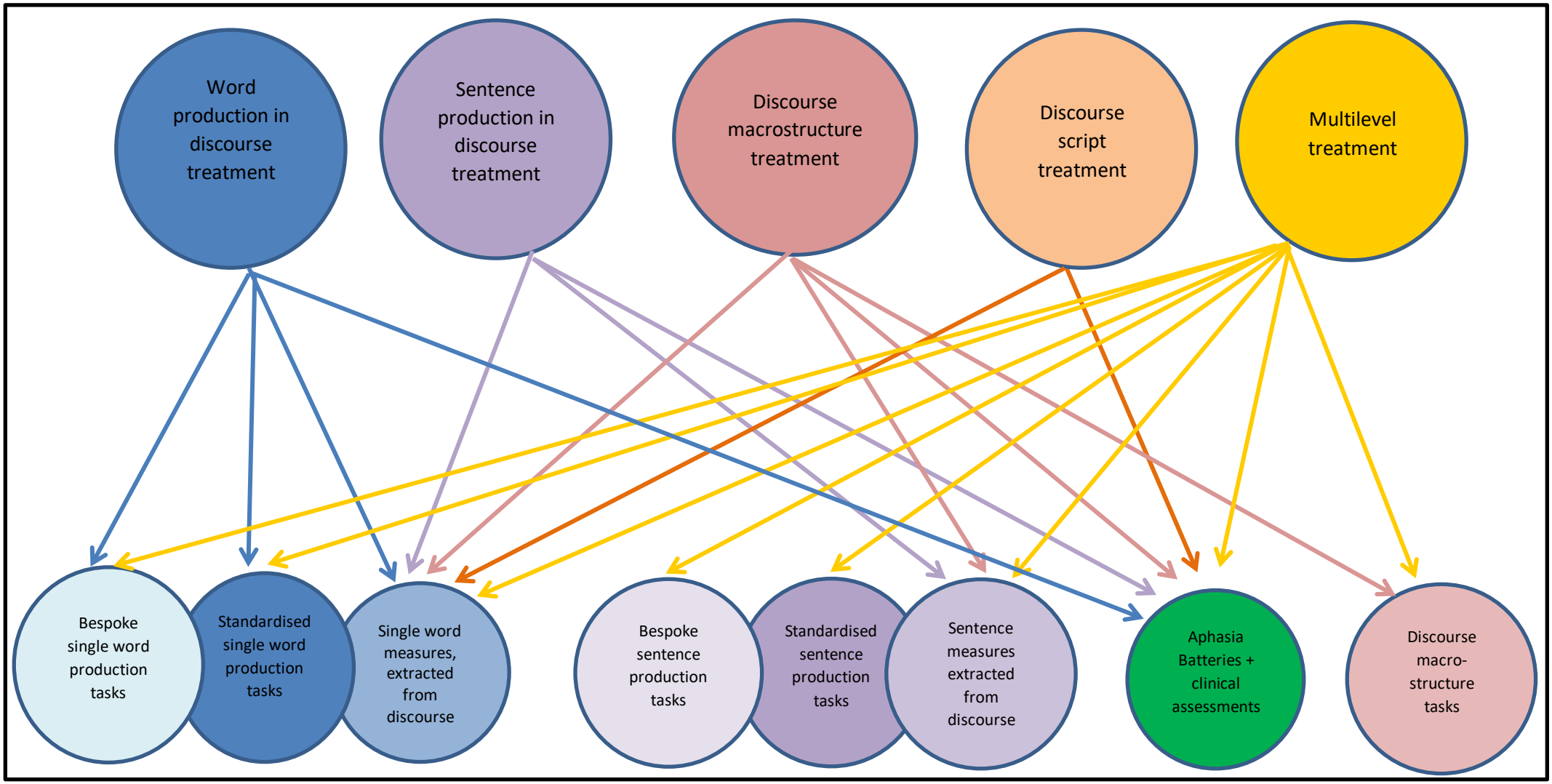


Figure 2

Visual representation of the range of treatment targets and associated outcome measures found in the reviewed papers



Appendix 1

PEDro-P and ROBIN-T and Scale items and scoring

Scale	Study designs assessed	Items
PEDro-P	Randomised & non-randomised control trials	1. Eligibility criteria were specified 2. Subjects randomly allocated to interventions 3.Allocation concealed 4. the intervention groups were similar at baseline regarding the most important prognostic indicators 5. Blinding of all subjects 6. Blinding of all therapists who administered 7. There was blinding of all assessors who measured at least one key outcome 8. Measures of at least one key outcome were obtained from more than 85% of the subjects initially allocated to groups

		9. All subjects for whom outcome measures were available received the treatment or control condition as allocated
		10. The results of between-intervention group statistical comparisons are reported for at least one key outcome
		11. The study provides both point measures and measures of variability for at least one key outcome
RoBIN-T	Single case	1. Design with control
		2. Randomisation
		3. Sampling of behaviour
		4. Blinding of people involved in the intervention
		5. Blinding of assessors
		6. Interrater agreement
		7. Treatment adherence
		8. Baseline characteristics

9. Setting

10. Dependent Variables

11. Independent variables

12. Raw data record

13. Data analysis

14. Replication

15. Generalisation

Appendix 2

12 item TIDieR Checklist (from Hoffman et al., 2014)

Brief name	Provide the name or a phrase that describes the intervention
Why	Describe any rationale, theory, or goal of the elements essential to the intervention
What	Materials: Describe any physical or informational materials used in the intervention, including those provided to participants or used in intervention delivery or in training of intervention providers. Provide information on where the materials can be accessed (such as online appendix, URL)
What	Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities
Who provided	For each category of intervention provider (such as psychologist, nursing assistant), describe their expertise, background, and any specific training given
How	Describe the modes of delivery (such as face to face or by some other mechanism, such as internet or telephone) of the intervention and whether it was provided individually or in a group
Where	Describe the type(s) of location(s) where the intervention occurred, including any necessary infrastructure or relevant features
When and how much	Describe the number of times the intervention was delivered and over what period of time including the number of sessions, their schedule, and their duration, intensity, or dose

Tailoring	If the intervention was planned to be personalised, titrated or adapted, then describe what, why, when, and how
Modifications	If the intervention was modified during the course of the study, describe the changes (what, why, when, and how)
How well	Planned: If intervention adherence or fidelity was assessed, describe how and by whom, and if any strategies were used to maintain or improve fidelity, describe them
How well	Actual: If intervention adherence or fidelity was assessed, describe the extent to which the intervention was delivered as planned

Appendix 3

Table showing, participant numbers, study design, and results reported in each group of studies

Word production in discourse					
Study	Measures	Change shown	Maintenance	Statistics	Design (level of evidence)
Antonnucci (2009)	Word production extracted from discourse	P2 improved on number of words, number of CIUs and % nouns; P3 improved on % CIUs; CIUs/minute and % nouns	Gains were stable at 6 weeks follow up	Effect sizes	Case series, N = 3 Only 2 participants reported owing to missed therapy sessions with P1
	Single word errors extracted from discourse	Yes	Yes	Descriptive data only	
	Standardised naming tool	No	N/A	Descriptive data only	
	Aphasia battery (WAB; Kertesz, 1982) *Also error changes: more semantic errors and semantic paraphasias suggestive of enhanced access to semantic knowledge (P2); higher levels of self-correction (P2 and P3)*	No	N/A	Descriptive data only	

Falconer & Antonnucci (2009)	Word production extracted from discourse	P1 and P 2 improved on CIU/min; P3 improved on % CIUs and % verbs; P4 improved on number of CIUs	Gains were stable at 6 weeks follow up for all bar P4	Effect sizes	Case series, N = 4
	Single word errors extracted from discourse	Yes	Yes	Descriptive data only	
	Standardised naming tool	P4 only	N/A	Descriptive data only	
	Aphasia battery (WAB; Kertesz, 1982) *Also error changes: decrease in deletions/non-specific terms; increase in semantic paraphasias (inconsistent across participants)*	None improved by clinically significant benchmark	N/A	Descriptive data only	
Gordon (2007)	Bespoke single word production tool	Yes (largely on treated items)	Gains maintained at 2 months follow up	Descriptive data only on all measures	Case series, N = 2
	Word production extracted from discourse	No			
	Aphasia battery (WAB; Kertesz, 1982)	Both participants improved by at least 10 points	Gains maintained at 2 months follow up		

Greenwood et al. (2010)	Bespoke single word production tool	Significant gains	Gains maintained 8 weeks post therapy	Testing of significance with McNemar and Wilcoxon matched samples tests	Single case design, N = 1
	Word production extracted from discourse	Increase in content words per conversational turn	Gains maintained 8 weeks post therapy	Descriptive data only	
	Single word errors extracted from discourse	Reduction in number of errors/content words	Gains maintained 8 weeks post therapy	Descriptive data only	
	Clinical assessment: (CDP; Swinburn & Byng, 2006)	Positive changes in ratings of activity, participation and emotional consequences	Gains maintained 8 weeks post therapy		
Nickels et al. (2016)	Bespoke single word production tool	Improvement could not be attributed to therapy	N/A	Weighted statistics (WEST-ROC) to examine whether change occurs over the treatment period	Case series, N = 4 Only 3 participants reported as one withdrew early in the study
	Word production extracted from discourse	No	N/A	Descriptive data only	

Sentence production in discourse					
Study	Measures	Change shown	Maintenance	Statistics	Design (level of evidence)
Cherney (2010a)	Aphasia Battery (WAB; Kertesz, 1982) Word production extracted from discourse	Improvement on WAB AQ (but not reading and writing sub scores) Improvement on words per minute and CIUs per minute Gains do not differ between computer and face to face delivery	Not assessed	Effect sizes to compare change over a treatment and no treatment period T-tests to compare outcomes between a group given computer ORLA and a group treated face to face	Delayed treatment design to compare treatment and no treatment period (N = 11) RCT to compare computer ORLA with face to face ORLA (N = 25)
Cherney (2010b) (same participants as above)	Aphasia Battery (WAB; Kertesz, 1982)	Whole group show change following therapy, after a stable baseline. The level of change (<5 points) is below clinical significance. All severity groups (mild, moderate and severe) showed change following treatment, but severe participants had an unstable baseline.	Not assessed	Effect sizes (Cohen's d)	Group study (N = 25) delayed treatment design comparing a no treatment period with a treatment period

	Clinical Assessment (Reading Comprehension Battery for Aphasia (RCBA); LaPointe & Horner, 1979) Word production extracted from discourse	Results for the whole group are not reported; there were no changes for the different severity groups. Results for the whole group are not reported; Some changes occurred on specific measures, but these are confined to the moderate/mild participants, 2 gains had medium effect sizes, 4 gains had small effect sizes; 3 changes on these measures occurred over the untreated period			
Goral & Kempler (2009)	Aphasia Battery (Boston Diagnostic Aphasia Examination (BDAE); Goodglass, Kaplan & Barresi, 2001) Aphasia Battery (Cognitive Linguistic Quick Test (CLQT); Helm-Estabrooks, 2001)	Only one subtest (auditory comprehension showed pre to post therapy improvement No change	Not assessed Not assessed	Descriptive data only Descriptive data only	Single case A-B-A-B design; A=treatment, B=no treatment

	Word and morphology production extracted from discourse	Change in overall productivity following second treatment block; increase in the % and variety of verbs following both treatment blocks; increase in verb morphology pre-post therapy	Changes in verb production were maintained 10 weeks after the 2nd treatment block	Effect sizes	
	Other: social-communicative rating scale devised for the study	Small but significant gain pre to post therapy	Not assessed	T-test to compare pre and post therapy scores	
Murray et al. (2007)	Bespoke sentence production tasks: probing of trained and untrained sentences in a trained (written) and untrained (spoken) modality	Improvement in the written production of trained sentence types; generalisation to untrained sentence types that were syntactically related to the trained types; no change in an untreated sentence type (passives) that was unrelated to the trained types; generalisation to spoken production with patterns that largely followed the written performance.	Improvements on sentence probes maintained 4 weeks after the end of therapy	Descriptive data only	Single case multiple baseline across behaviours design

	Word production extracted from discourse Sentence production extracted from discourse	Positive trend in most measures; greater improvement in spoken than written discourse Improvement not clear cut (e.g. no change in % grammatical utterances)	Maintained at 4-week follow up		
Wambaugh et al. (2013)	Word production extracted from discourse; CIU production and number of novel words	Increase in CIU production shown by all participants, but only in procedural discourse; CIU in personal discourse did not change, even during treatment phases that targeted personal discourse; improvements seen in the number of novel words produced in procedural discourse	Improvements in procedural discourse were maintained at 6 weeks follow up.	Conservative dual criterion method (number of probes above a mean and a trend line extrapolated from baseline data) Effect sizes	Case series, N = 3

Discourse macrostructure					
Study	Measures	Change shown	Maintenance	Statistics	Design/level of evidence
Carragher et al. (2015)	Word production extracted from discourse Macro structure from discourse	All participants communicated more salient ideas (content words) from two narratives. Changes were more evident in the simple than the complex narrative; there were changes in the number of salient ideas understood by conversation partners. The narrative sequence improved for all participants; changes were most evident in the simple narratives.	Not assessed	None employed; descriptive data only	Non-experimental case reports, N=4
Osiejuk (1991)	Word production extracted from discourse Sentence production extracted from discourse	Increased word production Increase in the number of clauses; but no reduction in the number of grammatical errors	Not assessed	None employed; reporting is very descriptive, some of the reported changes lack corroborating data	Single case report, no control

	Macro structure from discourse Aphasia Battery (selected tests from BDAE) Other: Wechsler Adult Intelligence Scale (WAIS); Wechsler, 2008)	Increased production of superstructure elements Improvements in auditory comprehension and oral expression Most subtests were unchanged; improvement noted on the Similarity test.			

Discourse scripts					
Study	Measures	Change shown	Maintenance	Statistics	Design/level of evidence
Cherney & Halper (2008)	Word production extracted from discourse	2 (/3) participants made gains on 2 scrips (/3); gains involved increased production of script related words, increased number of morphemes (for one participant) and increased rate of production	Not assessed	None employed; descriptive data only	Non-experimental single case reports, N =3
	Aphasia Battery (WAB; Kertesz, 1982)	One participant improved (/3)	Gains for one participant maintained at 6 weeks follow up		
	Clinical Assessment Quality of Communication Life (QCL); Paul et al., 2004)	One participant (/3) improved	Gains for one participant maintained at 6 weeks follow up		
	Clinical Assessment Communication Activities of Daily Living (CADL); Holland, 1980)	No participants improved	Gains for one participant maintained at 6 weeks follow up		

	Clinical Assessment Communication Effectiveness Index (CETI); Lomas et al., 1989	One participant (/3) improved			
Cherney et al. (2008)	Word and morphology production extracted from discourse Aphasia Battery (WAB; Kertesz, 1982) Clinical assessment (QCL; Paul et al. 2004) Clinical assessment (CADL; Holland, 1980)	All participants produced more script related words and morphemes, and improved in their rate of production 2 participants (/3) improved on the Aphasia Quotient One participant (/3) improved No participants improved	Not assessed	None used; descriptive statistics only	Non-experimental case reports N = 3
Lee et al. (2009)	Word production extracted from discourse	Across all participants (N=17) there was a mean improvement in the number of script related words produced (% change = 45.7) and mean improvement in the rate of production (% change = 137.5)	Not assessed	Statistics not used to evaluate the change (correlations were conducted to explore relationships, e.g. between the amount of practice and the % change)	Non-experimental control group study N = 17

Multi-level studies – 2 levels of language					
Study	Measures	Change shown	Maintenance	Statistics	Design/level of evidence
Carlomagno et al. (1991)	Word production extracted from discourse: Cookie Theft from BDAE; PACE message task Story retell	No change in content units/speaking rate on Cookie Theft More efficient transmission of messages in PACE task (reduced time and reduced prompts) Increased information units in story retell	Not assessed	Non-parametric comparisons of pre/post therapy scores	Group study (N=8) comparing pre and post therapy scores No experimental control
Hickin et al. (2015)	Aphasia Battery (Comprehensive Aphasia Test (CAT); Swinburn et al., 2005) Word production extracted from discourse Sentence production extracted from discourse	Improvement on spoken picture description from CAT (but not written) Increased word production More complete sentences and embeddings; more compound sentences; reduction in simple	Not assessed	Not all changes were evaluated statistically; sentence level changes in discourse were evaluated with chi square	Single case report (N=1); Double baseline - repeated administration of CAT pre therapy showed no change (n.b. many scores were within normal limits)

		unconnected sentences			
	Macro structure from discourse	No change in measures of coherence and story grammar			
	Other: Quality of life questions	More positive responses post therapy, e.g. about ability to contribute to conversations			
Marini et al. (2007)	Aphasia Battery: (Aachen Aphasia Test) & Clinical Assessment (CADL; Holland, 1980)	1 participant showed 'slightly better post therapy profile' on AAT and improvement on CADL; other participants showed no change	Discourse evaluated 3 weeks after therapy with Cookie Theft picture description; CIU scores are reported as greater than pre therapy, but comparative data are not given	Pre/post therapy scores compared with ANOVA Statistics not applied to maintenance scores	Group study (N = 3) no control
	Word production extracted from discourse	Increase in number of words (nouns and function words) and in speech rate; increase in number, % and rate of CIUs; increase in number of accurate and complete CIUs			
	Sentence production extracted from discourse	Increase in Mean Length of Utterance			

	Other: Ratings by naïve listeners of communicative exchanges	but not in the number of complete sentences Improved ratings following therapy			
Milman et al. (2014)	Aphasia Battery: (WAB; Kertesz, 1982) Standardised naming tool: Boston Naming Test (BNT); Kaplan, Goodglass & Weintraub, 1983) Bespoke single word (adjective) production tool Bespoke sentence production task Word production extracted from discourse Other: Cognitive measures	2 participants (/3) improved on AQ 2 participants (/3) improved 2 participants (/3) improved 2 participants (/3) improved 2 participants (/3) improved 1 participant improved on SCCAN	Probes administered 1 week and 1 month after the end of therapy; gains on trained words/ sentences maintained by the 2 participants who benefited from therapy	Effect sizes for probe results but most data are descriptive	Case series (N = 3)

Penn & Beecham (1992)	Macro structure from discourse Other: ratings of compensatory strategies employed in discourse	Improved narrative organisation and other macro features such as turn making Improved use of strategies across languages (despite therapy only being conducted in English)	Not assessed	Not employed; descriptive data only	Single case report (N =1); no experimental control

Multi-level studies – 3 levels of language					
Study	Measures	Change shown	Maintenance	Statistics	Design/level of evidence
Hoover et al. (2015)	Standardised naming tools: Northwestern verb naming test Philadelphia naming test	Improved naming of nouns and verbs post therapy	Gains maintained at one month follow up	All changes evaluated statistically	Repeated measures group design (N = 12)
	Standardised sentence production tool	Improved sentence production post therapy	Scores did not fall significantly post therapy; but scores returned to a level that was close to the second baseline		
	Word production extracted from discourse	No significant change in CIU or verb production			
	Sentence production extracted from discourse	Significant increase in the number of complete sentences following therapy	Gains maintained at one month follow up		
	Clinical Assessment: Assessment for Living with Aphasia (ALA) Kagan et al., 2011) and American Speech-Language-Hearing	Significant post therapy gains on ALA and ASHA FACS	Not assessed		

	Association Functional Assessment of Communication Skills for Adults (ASHA FACS); Frattali et al., 1995)				
Whitworth (2010)	Single word production extracted from discourse Sentence production extracted from discourse Macro structure from discourse	1 participant (/2) produced more verbs in discourse post therapy Both participants produced more 2/3 argument sentences post therapy (5 other sentence measures did not change) Both participants improved in the story grammar score	Gains are reported to be maintained 4 weeks post therapy, but data are not reported	Changes are shown to be statistically significant	Single case reports (N = 2); non-experimental
Whitworth et al. (2015)	Single word production extracted from discourse Sentence production extracted from discourse	No between group differences. In the within group analysis the NARNIA group produced more verbs post therapy No between group differences. Within group analysis showed that NARNIA group produced more 2/3	Verb changes maintained at 5week follow up Changes for NARNIA group maintained	Between and within group differences evaluated statistically	Pilot RCT (N = 14) comparing discourse intervention (NARNIA) with Usual Care

		argument structures post therapy; both groups produced more embedded sentences post therapy			
	Macro structure from discourse	Between group analysis showed more gains in orientation elements for NARNIA group; Gains in orientation also evident for NARNIA group in the within group analysis; more Body elements produced by Usual Care group post therapy	Orientation changes not maintained for NARNIA group; changes in Body maintained for Usual Care group		
	Standardised naming tools	No between group effects; within group analysis show both groups improve in noun production; NARNIA group improve on verb production	Not reported		
	Standardised sentence production tools	No significant gains			

No consensus					
Study	Measures	Change shown	Maintenance	Statistics	Design/level of evidence
Dietz et al. (2018)	Word production extracted from discourse Sentence production extracted from discourse Macro structure from discourse Aphasia Battery (WAB; Kertesz, 1982)	Clinically significant changes favouring AAC group in % words and CIUs; both groups show change in rate of CIU production Clinically significant increase in T-units (grammatically structured units of discourse) favouring AAC group Reduction of mazes (extraneous language) favouring AAC group Both groups improve in Aphasia Quotient (AAC > Usual Care)	Not assessed	T-test comparisons between groups and effect sizes computed for change scores	RCT (N = 12) comparing AAC therapy with Usual Care; pre and post therapy comparisons
McCall et al. (2009)	Word and morphology production extracted from discourse	Improvement in use of obligatory inflections and determiners, and in the proportion of closed class words; proportion of pronouns/nouns	Not assessed	Chi-square comparisons to determine significance of change	Single case (N =1) multiple baseline design; results from two episodes of treatment

	Sentence production extracted from discourse	improved following treatment 2 Improvements in proportion of words used in sentences, mean sentence length, and proportion of well-formed sentences; increased use of embeddings following treatment 2			

Supplemental File 1

Participant profiles, where reported.

Article	***Profile description***	***Language assessment***
Word production in discourse studies		
Antonucci (2009)	P2 Conduction aphasia WAB AQ 63 P3 Anomic aphasia WAB AQ 90.2	WAB AQ
Falconer & Antonucci (2012)	P1 Conduction aphasia WAB-R AQ 69.6 (multiple CVAs) P2 Conduction aphasia WAB-R AQ 61 (single CVA) P3 Broca's aphasia WAB-R AQ 33.6 (TBI) P4 Transcortical-motor aphasia WAB-R AQ 52.4 (multiple CVAs)	WAB-R AQ
Gordon (2007)	2 PWA: one 14 months TPO with moderate-severe Broca's aphasia (BDAE severity 1-2); one 9 years post-head-injury with Broca's aphasia (BDAE rating 3) with apraxia	BDAE Severity Rating Scale and subtests

Greenwood et al. (2010)	1 PWA with anomia	Pyramids and Palm Trees Extensive background assessment reported in Table 1, although names of assessment(s) are not reported. Some assumed to be PALPA.
Nickels et al. (2016)	3 PWA, strokes 4, 5 and 21 years prior; 2 mild non-fluent and 1 moderate conduction	Comprehensive Aphasia Test
Sentence production in discourse studies		
Cherney (2010a)	25 PWA with chronic non-fluent but not global aphasia Overall WAB AQ mean=53.74, SD 25.34, range 9.7 to 81.5 WAB AQ Severe n=6, mean=13.73, SD 3.97, range 9.7 to 21.1 WAB AQ Moderate n=9, mean=54.74, SD 3.53, range 49.4 to 60.7 WAB AQ Mild n=10, mean=76.84, SD 5.56, range 66 to 81.5	WAB AQ

Cherney (2010b)	N=11 PWA in computer ORLA WAB AQ mean 62 (same PWA as 2010a)) N=14 PWA in SLP ORLA WAB AQ mean 47.3 (same PWA as 2010a)	WAB AQ
Goral & Kempler (2009)	1 PWA, 12 years TPO, chronic non-fluent with moderate difficulty	BDAE
Murray et al. (2007)	1 PWA, 2 years TPO, described with moderate severe Broca's aphasia, with WAB AQ 68	WAB AQ
Wambaugh et al. (2013)	3 PWA, 12-424 months TPO, described as mild and mild to moderately aphasic, with WAB AQ 74.1 (anomic), 73.8 (Broca's), and 91.8 (anomic)	WAB AQ
Discourse macrostructure studies		
Carragher et al. (2015)	4 PWA, 26-80 months TPO, aged 38-70 years	BNT and OANB verb naming More detail provided in Carragher et al. (2013) paper
Osiejuk (1991)	1 PWA, 8 months TPO, moderate anterior aphasia	BDAE

Discourse script studies		
Cherney et al. (2008)	3 PWA, aged 65-78 years; one participant with moderate severe Broca's; moderate Wernicke's; and moderate anomic. WAB AQ 50.4, 73.2, and 62.8	WAB AQ
Cherney & Halper (2008)	3 PWA, aged 50-75 years; non-fluent and fluent. WAB AQ of 51.4, 51.9, and 81	WAB AQ
Lee et al. (2009)	17 PWA, aged 31-70 years; mean of 66 months TPO (range 11-274 months). Mean WAB AQ 65.1, range 30.5-85.3	WAB AQ
Multi-level studies - 2 levels		
Carlomagno et al. (1991)	8 PWA at least 8 months TPO (details for age, gender, education, TPO not reported). Described as having Broca's, anomic, Wernicke's and moderate global aphasia; not test data reported.	Clinical judgment
Hickin et al. (2015)	1 PWA, 24 years, 2 years TPO, mild aphasia, Language Battery of the CAT	CAT
Marini et al. (2007)	3 PWA, aged 44-53 years; 12, 18 and 22weeks TPO, described as having moderate non-fluent aphasia	Aachen Aphasia Test (Italian version)

Milman et al. (2014)	3 PWA, aged 56-68 years, 1-7 years TPO, described as having Broca's aphasia on WAB-R, with WAB AQ of 35.8, 42.2, and 53.1	WAB-R AQ
Penn & Beecham (1992)	1 PWA, aged 38 years, mildly aphasic, WAB AQ 77.1; WAB translated into Afrikaans, Zulu, and Pedi and administered 4 times. TPO not clear.	WAB AQ
Multi-level studies – 3 levels		
Hoover et al. (2015)	12 PWA, single L CVA, 8.75 years TPO (range 1.5 -20 years), described as non-fluent profile of aphasia (Broca's)	Clinical assessment of non-fluent profile of aphasia & difficulty with verb and sentence production
Whitworth (2010)	2 PWA, 41-62 years, 7-27 months TPO, fluent and non-fluent	The Northwestern Test of Verbs and Sentences; Test of Thematic and Syntactic Structure
Whitworth et al. (2015)	8 PWA in NARNIA, aged 42-87 years, mean of 21 months TPO (range 2-49 months), described as mild-moderate aphasia	WAB-R

No consensus		
Dietz et al. (2018).	6 PWA, WAB-R AQs of 40.8 (global) 68.6 (Broca's), 48.9 (Broca's), 71.2 (conduction), 37.6 (Wernicke's), and 82.4 (anomic)	WAB-R AQ
McCall et al. (2009)	1 PWA, 59 years, 24 months TPO, WAB AQ of 64.8 (Broca's)	WAB AQ

Supplemental File 2

Dosage and intensity of intervention programmes

	Duration (weeks)	**# sessions per week**	**Total dose (hours) where calculable**
Words in discourse	6 – 16	1 – 2	9 – 16
Sentences in discourse	4 – 22	2 – 5	10 – 45
Discourse macrostructure	6-10	1 – 3	9 – 45
Discourse scripts	9	5 - 7	9 hours 9 minutes – 68.61 hours[6]
Multi-level (2 levels)	-	<1 to 5	14 – 70
Multi-level (3 levels)	5-10	2 - 4.5	-
No consensus	4-12	2.91 – 3	12 – 35

[6] Lee et al. removed participant COLPA as an outlier, as she exceeded the median treatment time by over six times over the interquartile range (total treatment = 151.85 hours, mean hours per week = 16.87 hours).