Aphasiology

ISSN: 0268-7038 (Print) 1464-5041 (Online) Journal homepage: www.tandfonline.com/journals/paph20

Reconstructing sentence processing in aphasia: a randomised control trial of a usage-based intervention

Claudia Bruns, Fern Rodgers, Kerry Dathan, Michael Dean, Jane Warren, Victoria Fleming, Amir-Homayoun Javadi & Rosemary Varley

To cite this article: Claudia Bruns, Fern Rodgers, Kerry Dathan, Michael Dean, Jane Warren, Victoria Fleming, Amir-Homayoun Javadi & Rosemary Varley (11 Jul 2025): Reconstructing sentence processing in aphasia: a randomised control trial of a usage-based intervention, Aphasiology, DOI: 10.1080/02687038.2025.2530572

To link to this article: https://doi.org/10.1080/02687038.2025.2530572

APHASIOLOGY
https://doi.org/10.1080/02687038.2025.2530572

Routledge
Taylor & Francis Group

Reconstructing sentence processing in aphasia: a randomised control trial of a usage-based intervention

Claudia Bruns[a], Fern Rodgers[a], Kerry Dathan[a], Michael Dean[a], Jane Warren[a], Victoria Fleming[a], Amir-Homayoun Javadi[a,b,3] and Rosemary Varley[a]

[a]Language & Cognition, Psychology & Language Sciences, University College London, London, UK; [b]School of Psychology, University of Kent, Canterbury, UK; [3]School of Rehabilitation, Tehran University of Medical Sciences, Tehran, Iran

ABSTRACT

Background: Listening to and producing sentences is a cornerstone of typical language exchanges. Therapy for aphasic impairments has tended to focus on single-word processing, with comparatively few sentence-level therapies. Usage-based Construction Grammar is an approach to language in which frequency of use of grammatical constructions plays a central role in representation and processing of structures. We report findings from a usage-based sentence intervention: UTILISE (Unification Therapy Integrating LexIcon and SEntences). The intervention began by priming high-frequency constructions (e.g. *I like it*) via listening tasks and then practice of production. Subsequently, different lexical items were inserted to slots around the verb (e.g. *I like coffee now*) to increase communicative options.

Aims: To evaluate the impact of UTILISE on participants' spoken sentence production and comprehension abilities.

Methods & procedures: Participants with chronic aphasia ($n = 39$) were recruited to a two-arm randomised control trial, with 33 participants completing the intervention. At trial entry, participants were randomised to Immediate/Deferred conditions, allowing for treatment/no treatment comparison. Two baseline measures were taken (four-week interval in the Immediate condition; eight-week interval in Deferred). A four-week therapy phase comprised two auditory processing tasks and one spoken sentence production task, delivered over 12 in-person sessions. Outcomes were measured immediately post-intervention and after an eight-week maintenance phase. Main outcome measures were: sentence production in narratives, measured as ratio of three-word combinations to total words in connected speech (Connectivity); spoken sentence comprehension (TROG-2) and quality of life (QoL) perceptions (SAQOL-39). Intervention acceptability was also evaluated, together with an untreated control task.

Outcomes & results: A between-group comparison of Connectivity and TROG-2 scores revealed no significant difference; however,

KEYWORDS

Aphasia; sentence comprehension; sentence production; therapy; randomised control trial; usage-based Construction Grammar

CONTACT Claudia Bruns c.bruns@ucl.ac.uk Language & Cognition, Psychology & Language Sciences, University College London, 2 Wakefield Street, Chandler House, London, WC1N 1PF, UK

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02687038.2025.2530572

when data were pooled across groups, linear mixed-effects models revealed gains following therapy in Connectivity, whereas increases in sentence comprehension (TROG-2) scores might be due to repeated exposure to the test. QoL perceptions improved, reaching significance on the SAQOL-39 communication sub-scale. Participants found the UTILISE intervention acceptable.
Conclusions: The study produced initial indications of the value of a usage-based sentence therapy, with increases in three-word combinations in connected speech, enhanced QoL ratings regarding communication, and high acceptability to participants. A number of factors may affect results: intervention was delivered at low-dose, and both production and comprehension measures represented distal measures. However, given these preliminary results, UTILISE has now been developed into an app enabling higher-dose intervention, and is currently under evaluation.
Trial registration: Prospectively registered on 13/09/2019 at ISRCTN14466044

1. Introduction

Aphasia is not a rare condition as it is present in 21–38% of individuals who experience a stroke (Berthier, 2005). It has profound consequences for the individual and their communication partners, limiting the capacity to exchange thoughts and feelings, access print and digital media, and participate in work and lifelong education (Lam & Wodchis, 2010; Ross & Wertz, 2003). Aphasic impairment traditionally has been addressed through the dichotomy of lexicon and grammar, based on the observation that some individuals appear to have disproportionate disruption in one or other of these language subdomains. However, most people with aphasia have both vocabulary and grammatical impairments that often encompass comprehension and production to varying degrees (Wilshire et al., 2014).

Aphasia interventions have tended to target single-word processing, with therapies predominantly directed to production of nouns. Other studies target verb naming (Edmonds, 2016; Marshall et al., 2018) or, more recently, a wider range of word classes (Upton et al., 2024). A typical result for noun interventions is improvement of treated words, but limited generalisation to untreated items or spontaneous speech (e.g., Palmer et al., 2019). However, higher-dose therapy (Upton et al., 2024) or speeded retrieval (Conroy et al., 2018), appear to achieve some transfer beyond picture naming, and therapies targeting verbs and their associated arguments are reported to have greater potential for generalisation to connected speech (Edmonds, 2016; Hickin et al., 2019, 2022; Webster et al., 2015).

Despite the development of effective naming therapies, there has been less innovation in the domain of sentence processing. A PubMed literature search conducted in November 2024 revealed that, since 2007, there were 159 trials of lexical therapies versus only 31 sentence-level studies.[1] This comparative scarcity of sentence-level interventions potentially limits the functional outcomes of aphasia therapies as typical language exchanges involve sentences. The sentence interventions that have been reported often focus on structures that rarely occur in day-to-day communication, such as passives

or object relatives, although there is some evidence that stimulation of complex forms can result in generalisation to simpler structures (Thompson et al., 2003, 2010). Other sentence therapies focus on verbs (e.g., VNeST; Edmonds et al., 2009), with stimulation of elements of sentence structure around a verb. High-imageability verbs such as *measure* and *chop* are at the core of these therapies, however they are of relatively low-frequency due to a trade-off between semantic specificity and usage frequency. Typically, evidence levels for sentence therapies would be classed as early stage, with predominantly small-scale studies employing case series designs, and little longer-term follow-up.

The impact of frequency at a single-word level is well established in aphasia, as well as in neurotypical language processing (Brysbaert & Ellis, 2016; Kittredge et al., 2008; Nickels & Howard, 1995). Higher-frequency words are processed more rapidly by neurotypical speakers and are more resilient to disruption in aphasia. Frequency also modulates processing at a multi-word level as some word strings are more frequent than others. For example, in the British National Corpus (BNC, 2007), the string *I don't know her* (raw frequency = 26) is much more frequent than *The girl chased the boy* (zero occurrences in the BNC), despite the former having more complex verb phrase structure. String frequency confers processing advantages in both neurotypical processing and pathology (Conklin & Schmitt, 2012; Siyanova-Chanturia et al., 2011; Wray, 2002). For example, Zimmerer et al. (2016, 2018, 2020) reveal retention of higher-frequency strings in the spontaneous speech of people with post-stroke aphasia, Alzheimer's dementia and primary progressive aphasia. Bruns et al. (2019) detail retention and appropriate use of the high-frequency string *I don't know* in speakers with non-fluent aphasia.

In this report, we describe an intervention study directed at sentence processing impairments in aphasia. The intervention (UTILISE – Unification Therapy Integrating LexIcon and SEntences) targets both spoken sentence comprehension and production. The content of UTILISE is informed by Construction Grammar (CxG; Croft, 2001; Dąbrowska, 2014; Goldberg, 1995, 2006) and, in particular, a usage-based variant of CxG. In contrast to generative models of grammar (e.g., Chomsky, 1981), CxG proposes that a speaker's linguistic knowledge is captured in terms of constructions. Constructions are linguistic units of varying size and degree of abstractness, stored in long-term memory. Some constructions are small and concrete (i.e., their phonological/orthographic form is specified) such as single words, but larger, concrete units can also be stored, including idioms and formulas such as *I don't know*. Similarly, abstract constructions may be small, such as [NOUN] or [VERB], or larger, such as [X VERBS Y]. Some constructions are more flexible, containing fixed elements plus open slots into which other constructions can be inserted (e.g., *Where's* ____). Frequency of use is one factor influencing whether a construction is stored in concrete form, irrespective of construction size. In the case of large concrete constructions, the frequent co-occurrence of sub-components of the string leads to associative links between sub-components, resulting in facilitation of access from one component to the next. In some cases, the sub-elements become fused and stored as a chunk (Bello-Lepe et al., 2024; Gahl, 2003, 2008; Whiteside & Varley, 1998). High usage frequency reflects high functional value, and these more concrete constructions are used to manage everyday interactions and express common communicative intentions.

A first step in developing UTILISE was to identify twenty fixed or semi-fixed constructions (i.e., formulaic frames) from the BNC of high frequency and that supported a range of communicative functions (Appendix 1). These constructions formed the core content of

the therapy. Because of their high frequency, these constructions are more likely to be resilient to disruption and/or more easily stimulable in therapy. The verbs treated in UTILISE are of high frequency and are adaptable to varying semantic contexts (e.g., *made, feel, want*). From these core constructions, we aimed to increase the communicative repertoire of the person with aphasia (PwA) through insertion of different lexical items, including high-frequency pronouns, into positions around the verb, or by adding additional elements such as an adjunct. In this way, the therapy develops from, for example, *I don't know* to *We don't know that now*. Other aphasia therapies have also employed formulaic frames. For example, later iterations of intensive language-action therapy (Difrancesco et al., 2012) describe use of interactional formulas such as *Can I have X* in a card exchange game. Similarly, formulas are likely to appear in script therapies for use in specific situations (e.g., Kaye & Cherney, 2016). Melodic intonation therapy also stimulates common/highly functional sentence frames (Helm-Estabrooks et al., 1989). UTILISE differs from these approaches: first, in its explicit basis in usage-based CxG and frequency values from the BNC, and second, in the stimulation and subsequent systematic extension of formulas, facilitating a range of exchanges.

The core constructions were embedded into a three-task therapy. The two initial steps involved listening to sentences, with the third step requiring spoken sentence production. The design of UTILISE was informed by insights from the existing cognitive neuroscience and neurorehabilitation literature. First, based on connectivity between input and output systems (Bolognini et al., 2016), the targeted constructions were primed in the listening tasks prior to any demand for production (Hartsuiker & Kolk, 1998; Lee & Man, 2017). Priming the core constructions also incorporated principles of errorless learning, or more properly, error-reduction, so as to increase the likelihood of fluent production in the speaking task (Fillingham et al., 2006; Whiteside et al., 2012). All tasks were computerised to enable close control of the intervention protocol and to explore potential for a future self-administered digital, higher-dose therapy (Berthier & Pulvermüller, 2011; Leff et al., 2021; Varley, 2011). Full description of the intervention can be found in Appendix 2, alongside a Template for Intervention Description and Replication (TIDieR) statement (Hoffmann et al., 2014).

In our original study plan (Varley et al., 2020), we intended to combine the behavioural therapy with non-invasive brain stimulation (tDCS) and MRI scanning to determine lesion size and location. Recruitment began in autumn 2019 but was interrupted by the COVID-19 pandemic. 12 participants were recruited prior to the pandemic, of whom three were randomised to active tDCS and two to sham stimulation conditions (the remaining seven did not fit criteria for MRI/tDCS and were allocated to behavioural therapy only). To permit rapid restart of recruitment post-pandemic, where we were not permitted to work within two metres of a participant, we eliminated the tDCS and MRI components of the study.

The study design post-pandemic was a two-arm randomised control trial (RCT), employing a waitlist control design (Varley et al., 2021). Participants were randomised to either Immediate or Deferred conditions. Participants completed two baseline evaluations: with a four-week gap in the Immediate condition and eight-weeks in the Deferred condition. The treatment was delivered in a university clinic, at an intensive frequency (12 sessions in a four-week period), but at relatively low-dose (session length of approximately 45–60 minutes with 40–45 minutes of time-on-task). Outcomes were measured at the end of the therapy phase and after an eight-week no-treatment period.

Our primary aim was to explore the efficacy of UTILISE on aphasic sentence comprehension and production impairments. The primary outcome measures were connectivity of spontaneous speech (personal narratives and retelling of a cartoon series) and spoken sentence comprehension, measured with the Test of Reception of Grammar (TROG-2; Bishop, 2003). Connectivity was operationalised as the ratio of grammatical three-word combinations (trigrams) to total number of words, measured via the Frequency in Language Analysis Tool (FLAT) automated software (Zimmerer et al., 2016, see Methods and Appendices for further detail). Both of these outcome measures are distal in that they did not specifically assess structures treated in therapy. Therefore, we also included a more proximal, although less functional, measure via a study-specific story completion test (Goodglass et al., 1972) that probed use of treated constructions and frequency- and length-matched untreated control constructions. We report the story completion test methods and results in an associated report (Bruns et al., in prep.). Perceptions of quality-of-life (QoL) before and after intervention were probed with the Stroke and Aphasia Quality of Life Scale (SAQOL-39g, Hilari et al., 2003). We also recorded acceptability of the intervention for participants. Finally, the impact of the therapy on an unrelated language behaviour was evaluated with a shortened form of the Written Synonym Matching Task from the Action for Dysphasic Adults Auditory Comprehension Battery (A.D.A., Franklin et al., 1992).

We hypothesised that UTILISE would be effective in enhancing connectivity in narrative speech production and improving sentence comprehension. We addressed the following research questions:

(1) Does UTILISE improve connected speech in comparison to a no-treatment waitlist control?
(2) Does intervention improve comprehension of spoken sentences?
(3) Are treatment effects, if any, maintained following withdrawal of intervention?
(4) What is the relationship between treatment outcome and baseline demographic and behavioural profiling variables?
(5) What is the impact of the intervention on:
    (a) perceptions of QoL?
    (b) an untreated control language behaviour (written synonym judgement)?
    (c) acceptability judgement?

2. Methods

The trial was prospectively registered at ISRCTN14466044, and protocols pre-registered at OSF (pre-pandemic: Varley et al., 2020; post-pandemic: Varley et al., 2021). Ethical approval was granted by the UCL Research Ethics Committee (8123/001). All recruits gave written, informed consent to participation in the study.

A revised power calculation for the modified post-COVID study (with two conditions: Immediate vs. Deferred therapy) was performed using Connectivity data from Bruns et al. (2021). The minimum sample size for reaching a medium effect size (d = .75) in a trial with a control group (Deferred trial entry) was 23 (one-tailed independent samples t-test; alpha level of .05; power of .8; using baseline average versus post-intervention trigram connectivity difference scores $M = .03$, $SD = .04$ as an indicator of the effect size). This

estimation is based on G*Power (version 3.1.9, Faul et al., 2009), where a hypothetical control group was assigned a mean of 0 and the same standard deviation as in Bruns et al. (2021). The recruitment target was therefore 46 participants; however, we ended recruitment before this target was reached due to delays resulting from the pandemic.

2.1. Participants

Volunteers were recruited via a university research register and community groups. Study inclusion/exclusion criteria were modified during the course of the study due to the impact of the COVID pandemic and change in protocol. Initial recruitment criteria included factors relating to MRI and tDCS safety. Final inclusion criteria were: adults (age ≥18 years) with capacity to consent; premorbid competence in English; no history of developmental speech/language disorder or progressive neurological disorder; moderate-to-severe aphasia after a single stroke (either or both of: < 16 blocks correct on TROG-2 (Bishop, 2003); spoken sentence production difficulties limiting output to incomplete and/or simple sentences). We did not restrict recruitment to participants with non-fluent/agrammatic aphasia and participants had varied behavioural and likely lesion profiles. Although we were able to characterise behaviour, the removal of structural imaging from the protocol resulted in inability to describe lesion site/extent. Recruits were in the chronic phase of recovery (> 6 months post-stroke) and had sufficient auditory and visual acuity to interact with a computer-based task.

Forty-seven potential recruits were assessed for eligibility and 39 (83%) met recruitment criteria and consented to participate in the study. These participants were randomised to Immediate ($n = 20$) or Deferred ($n = 19$) conditions. Block randomisation was performed via a computer-generated code from an external randomisation service (www.sealedenvelope.com) to ensure approximately equal numbers in each condition. Randomisation was stratified by sex, with variable block size that was not known to researchers (CB, FR, KD) responsible for assessment and intervention. Randomisation was managed by a researcher (RV) who was masked to case. Participants were allocated to Immediate vs. Deferred conditions after a first baseline assessment and at the first meeting assessors were masked to allocation. After this point assessors/therapists could not be masked to allocation or phase. Participants were not informed of the two-condition study design (although this information might be available to them via the online pre-registrations of the protocol study). Instead, they were advised that therapy would begin when the next space was available. In this way, we attempted to mask participants of their allocation to condition. The progress of participants through the study is shown in a CONSORT diagram (Schulz et al., 2010) (Figure 1). A check of allocation at the end of the study confirmed that all participants received their assigned treatment. One participant was excluded after randomisation when new diagnostic information emerged indicating this recruit was no longer eligible for participation. At the first outcome measure (O1) there were 17 participants in the Immediate condition, and 16 in Deferred. Reasons for drop-out, where given, are recorded in Figure 1. Participant numbers at the second outcome assessment (O2) were 14 in the Immediate condition and 13 in Deferred. Of the six participants lost to eight-week follow-up, four were due to interruption by the pandemic.

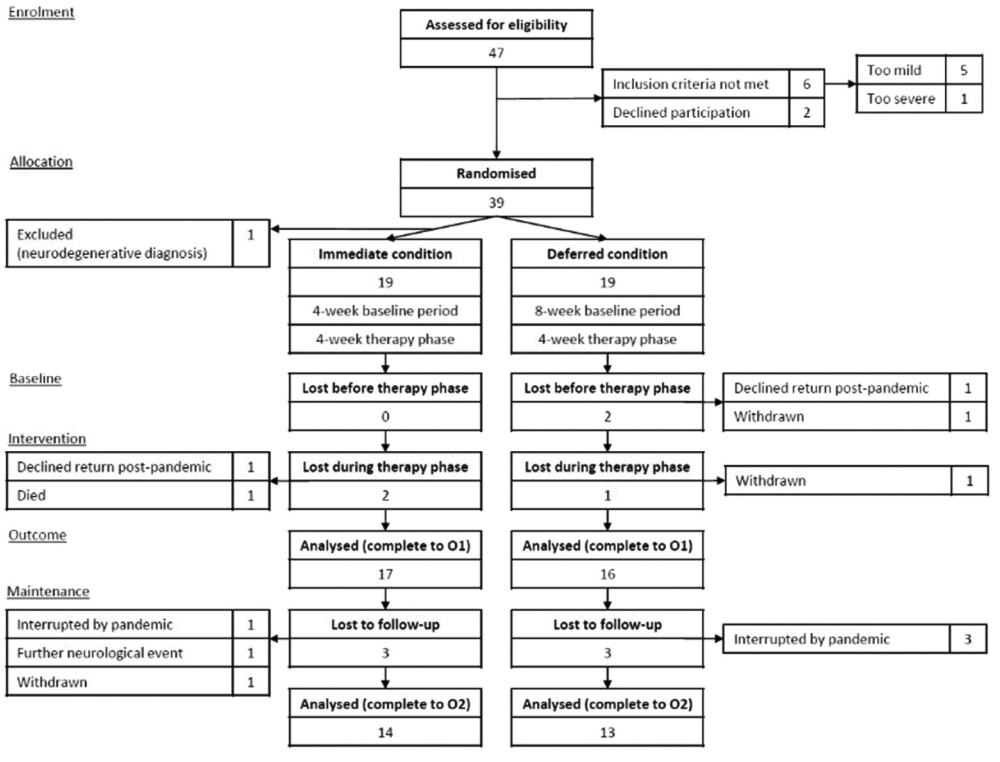

Figure 1. CONSORT (Schulz et al., 2010) flowchart of UTILISE study.

Demographic information was recorded (age, sex, previous employment, and years of education), as well as relevant medical history, including date of stroke. Employment/education was assigned an ordinal value based on Office of National Statistics coding (Standard Occupational Classification, 2020). Handedness was determined via a short-form version of the Edinburgh Handedness Inventory (EHI; adapted from Veale, 2014). Language and cognitive profiling measures were collected across the two baseline sessions (B1 and B2). Profiling data are presented in Table 1. Overall aphasia severity was evaluated via a composite score of the Boston Naming Test (BNT; Kaplan et al., 2001) and a spoken word-picture matching test (Comprehensive Aphasia Test (CAT) (Swinburn et al., 2004). Phonological working memory was assessed via a digit span task (Subtest 13 PALPA, Kay et al., 1992). Participants could respond either by repetition or pointing to digits on a number card. Non-linguistic cognition was evaluated with the Brixton Test (Burgess & Shallice, 1997), measuring executive function, while problem-solving in the visuo-spatial domain was assessed with the WASI-II Matrix Reasoning subtest (Wechsler, 2011).

We compared baseline profiles of the two groups using independent-sample t-tests or Mann-Whitney U test (reported by Wilcoxon's W). Table 1 shows that there were no significant differences between groups on demographic or other profiling measures, apart from the aphasia severity composite and the BNT score, where the Deferred group displayed more severe impairment (BNT: $W = 193.500$,

Table 1. Demographic and profiling data by group.

	Immediate (n = 17)	Deferred (n = 16)	Independent samples t-test/Mann-Whitney test	Effect size (Cohen's d/ rank-biserial correlation)
Age	57.41 (12.82)	63.94 (8.06)	W = 101.500, *p* = *.220*	d = .254
Sex	41% female (*n* = 7)	31% female (*n* = 5)		
Time post onset (in months)	66.06 (53.57)	102.69 (90.22)	W = 107.500, *p* = *.313*	r_B = .210
Years of education	14.82 (2.16)	15.5 (2.28)	W = 113.500, *p* = *.403*	r_B = .165
Occupation coding	4.47 (2.48)	3.44 (1.63)	W = 168.00, *p* = *.240*	r_B = .235
Handedness (EHI) [1] (61–100 = right handers)	88.24 (26.69)	82.81 (49.34)	W = 146.00, *p* = *.681*	r_B = .074
Object naming (BNT) (max. 60) [2]	22.88 (17.24)	10.56 (13.88)	W = 193.500, *p* = *.039*	r_B = .423
Spoken word comprehension (CAT) (max. 30) [3]	25.71 (4.51)	21.75 (6.56)	W = 182.500, *p* = *.095*	r_B = .342
Spatial anticipation executive functioning (Brixton Test) (max. 54) [4]	24.24 (8.58)	25.44 (14.2)	t*(31)* = 0.296, *p* = *.769*	d = .103
Digit span by repetition (PALPA-13) (max. 7) [5]	3.12 (1.54)	1.93 (1.75)	W = 169.500, *p* = *.103*	r_B = .329
Non-verbal IQ (WASI-II MR) (max. 30) [6]	13.41 (5.57)	14.25 (5.34)	t*(31)* = 0.441, *p* = *.662*	d = .154
Composite severity score*	61.91 (20.4)	45.05 (19.35)	t*(31)* = 2.433, *p* = *.021*	d = .847

Note. *Composite severity score was determined by averaging each individual's BNT and CAT spoken word comprehension percentage scores (a correlation analysis showed that both variables are statistically highly related: *r* = .62, *p* < .001). Values in round brackets indicate SD. [1] Edinburgh Handedness Inventory (EHI) short form adapted from (Veale, 2014); [2] Boston Naming Test (BNT; Kaplan et al., 2001); [3] Comprehensive Aphasia Test (CAT; Swinburn et al., 2004); [4] The Brixton Test (Burgess & Shallice, 1997); [5] Auditory Digit Repetition Span/digit pointing task: Subtest 13 from the Psycholinguistic Assessments of Language Processing in Aphasia (PALPA; Kay et al., 1992); [6] WASI-II Matrix Reasoning subtest: Wechsler Abbreviated Scale of Intelligence – Second Edition (Wechsler, 2011).

p = .039, r_B = .423, medium effect; composite severity score: t(31) = 2.433, *p* = .021, d = .847, large effect).

2.2. Procedure

Appendix 3 presents an overview of assessments and their timepoints in line with a SPIRIT template (Chan et al., 2013).

2.2.1. Primary outcome measures

Probes of Connectivity in spoken narratives (ratio of grammatical three-word combinations (trigrams) to total number of words) and TROG-2 were completed in baseline (B1 and B2) and outcome (O1 and O2) sessions. Repeated baselines allowed evaluation of stability of behaviour prior to intervention. Performance was assessed immediately after intervention (O1), and after an eight-week no-treatment period (O2) to assess maintenance of any treatment gains. The primary efficacy endpoint was the comparison between B2-O1 in the Immediate group versus B1-B2 for the Deferred group. Figure 2 (simulated data) illustrates how trigram ratios measure speech connectivity. The values are derived by automated software (FLAT,

Figure 2. Simulated data to illustrate trigram connectivity.

Zimmerer et al., 2016) which uses a "moving window" such that each grammatical trigram contributes to the final value (i.e., *a man phones; man phones his; his friend and*, etc.).

Two tasks were employed to elicit samples of personal narratives: an account of the participant's last holiday, and their activities at the previous weekend. "Last Weekend" was recorded at B1 and O2 and "Last Holiday" at B2 and O1. Narrations of two eight-picture cartoon series were also recorded (Jogging and Dinner Party, Fletcher & Birt, 1983). Jogging was recorded at B1 and O2, and Dinner Party at B2 and O1. Audio-recordings were allocated a code that masked the assessment timepoint and participant's allocation to Immediate/Deferred conditions. Orthographic transcription was performed by members of the research team who had no contact with participants. After transcription, samples were tagged with the conventions necessary for FLAT analysis (Zimmerer et al., 2016). FLAT generates a number of variables (Zimmerer et al., 2020) and for the purposes of this study, we report only trigram-based connectivity as our pre-registered primary outcome measure. A key tag was the separator (<.>) which is placed at utterance boundaries (e.g., *house* <. > *cooking* <. > *is a fish*) or between ungrammatical combinations (e.g., *really nice* <. > *them)*. Appendix 4 presents a sample FLAT output. Transcribers/taggers were trained in pairs using archive speech samples. They independently transcribed the samples and then compared their transcription to a master transcript. Errors were identified and the process repeated until there was a close correspondence between transcripts (e.g., one pair reached 97–98% agreement). A similar process was followed in training tagging (e.g., with the above pair achieving agreement of 95%). As a final check, one researcher (RV) sampled masked transcriptions against audio files and checked tagging before machine analysis was performed.

Comprehension of spoken sentences was measured with TROG-2 (Bishop, 2003). This sentence-picture matching task is organised by developmental sequence. Participants match a spoken sentence to one of four pictures. It probes a range of constructions, including structures that do not appear in many aphasia tests (e.g., pronouns, negation). Standard scoring criteria were used, with testing discontinued after five consecutive failed blocks (one or more errors per 4-item block). Administration of the test was video-recorded, and subsequently 11% of recordings were randomly sampled to allow fidelity checks on administration (e.g., adherence to the cut-off criterion, absence of cueing of responses) and scoring. These checks demonstrated high fidelity in both administration and scoring (99% and 100% agreement, respectively).

2.2.2. Secondary outcome measures

Perceptions of QoL before and after intervention were evaluated with the SAQOL-39 (Hilari et al., 2003). The assessment was performed once at baseline (B2) and after the no-treatment maintenance period (O2).

Performance on a shortened version of the A.D.A Written Synonym Judgement Test (Franklin et al., 1992) was also measured at all four timepoints to evaluate change in an unrelated language behaviour (written single word processing) to determine if there were any general effects of intervention. This measure acted as an untreated control measure and was not predicted to change following intervention. 80 items were selected from the test, with stimuli biased towards low-imageability ($n = 32$, out of 40) and low-frequency ($n = 32$, out of 40) words to avoid a ceiling effect, and matched sets of 40 items were tested at each timepoint. Participants judged whether two written words were similar or different in meaning. Acceptability of intervention was probed with a study-specific questionnaire.

2.2.3. Intervention phase

Therapy was delivered in 12 face-to-face sessions in a university clinic, each including approximately 40–45 minutes of time-on-task (overall average across 33 participants = 42 minutes), over a four-week period. The therapy consisted of three tasks. The high frequency constructions identified in the BNC (e.g., *I made it*) appeared in each task. In Task One, participants made same/different judgements on spoken word strings. Task Two involved a word-monitoring paradigm where participants pressed a response button as soon as they heard a target word. In both tasks, there was implicit focus on verbs with proportionally more differences on verbs in Task One and verb targets in Task Two. Task Three focused on sentence production. For each sentence ($N = 20$ constructions), participants listened to the spoken sentence, imagined saying the sentence in unison with the audio recording, and then repeated the sentence aloud (Varley et al., 2016). After completing the core sentence (e.g., *go home*), each construction was systematically loosened and lengthened by insertion of new lexical content in the slots around the verb (e.g., *I go to work*), or through the addition of adjuncts (e.g., *I go home now*). See Appendix 2 for detail on tasks and TIDieR statement. Treatment fidelity was addressed through use of a manualised procedure for administering tasks.

2.3. Analyses

We employed an intention-to-treat principle, including all participants in main analyses for whom repeated baseline data were available (Immediate and Deferred groups) and those in the Immediate group with a post-therapy outcome measure.[2]

Using SPSS (version 29, IBM Corp., 2023), we assessed data using Shapiro-Wilk tests for fit to parametric criteria. The analyses for Connectivity and TROG-2 included: First, the stability of behaviour between the two baseline evaluations, using a 2×2 mixed analysis of variance (ANOVA) with a between-participant factor of Group (Immediate: $n = 17$, Deferred: $n = 17$) and a within-participant factor of Time (B1, B2); and subsequently, examination of treatment versus no-treatment, using a 2×2 mixed ANOVA with a between-participant factor of Group (Immediate: $n = 17$, Deferred: $n = 17$), and a within-participant factor of Time (B1, B2 for Deferred participants or O1 for Immediate

participants. Connectivity comprised the average narrative score per participant per timepoint (cartoon narration and personal narratives).

Due to the underpowered nature of the main analyses, we also conducted post-hoc exploratory analyses regarding overall treatment effects for Connectivity and TROG-2 which comprised linear mixed-effects models (LMMs), using lmer function from the lme4 package in R (Bates et al., 2015; R core team, 2025). Linear mixed-effects models were chosen to account for the hierarchical structure of the data, where each participant contributed multiple data points per timepoint (Connectivity), and to allow for the inclusion of partial data sets. This preserved the richness of the data and sample size, maximising statistical power. Furthermore, individual differences in aphasia severity could be accounted for by including severity as a fixed effect, while variability in Connectivity across participants could be accounted for by incorporating a random intercept for participants.

This analysis utilised the complete data set (i.e., Immediate and Deferred participants, $N = 34$). Connectivity comprised two scores per participant per timepoint (cartoon narration and personal narratives); TROG-2 comprised one score per timepoint (number of blocks correct). All models included fixed effects of Time and Aphasia Severity (as measured by BNT) with two different parameterisations (continuous and categorical), to identify the combination that best modelled changes in Connectivity and TROG-2 (i.e., which combination offered the clearest insight into how these predictors relate to the outcomes). This resulted in four models each for Connectivity and TROG-2 to examine all possible combinations of parameterisations (see Appendix 5).

Random intercepts for individual participants were included in all models to capture between-subject variability in severity. For Time, predictors were 0, 1, 2, 3 (continuous) and B1, B2, O1, O2 (categorical). Performance on the BNT was used as a proxy for Aphasia Severity; predictors were BNT raw score (continuous) and BNT severity group (Mild, Moderate, Severe; categorical). To convert BNT into severity groups, scores were subjected to k-means cluster analysis with an a-priori three-cluster solution that resulted in the following groups which we assigned as follows: Mild ($n = 8$, $M = 41.13$); Moderate ($n = 9$, $M = 24.11$); Severe ($n = 17$, $M = 2.88$).

Model comparisons were planned using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to identify the most parsimonious model(s) that best explained Connectivity and TROG-2 performance. Following selection of final models, model diagnostics were conducted to check for homoscedasticity, and normality of residuals and random effects. Planned contrasts were used to examine changes in outcomes between discrete timepoints. Models were estimated using maximum likelihood (ML) in the comparison stage for valid comparisons, and final models were refitted using restricted maximum likelihood (REML) to provide unbiased estimates of variance components. All reported models converged using the default algorithm in lme4 package. The clubSandwich package in R was used for investigating fixed effects as it provides robust standard errors for mixed effects models, accounting for potential heteroscedasticity (Pustejovsky, 2023). The emmeans package in R was used for analysis of planned contrasts.

Finally, we analysed relationships between demographic and profiling characteristics and treatment outcome (calculated from pre- versus post-treatment change scores; O1-B1) using correlational analysis in SPSS.

Table 2. Summary of connectivity and TROG-2 average scores (SD) over time by group.

					Change scores		
	B1	*B2*	*O1*	*O2*	*Baseline phase B2-B1*	*Therapy phase O1-B1 [O1-B2]*	*Maintenance phase O2-O1*
Connectivity	.30 (.19)	.30 (.18)	.32 (.19)	.39 (.18)	.0 (.0)	.02 (.09) [.02 (.07)]	.01 (.07)
TROG-2 blocks	4.59 (3.48)	5.53 (4.17)	5.71 (4.73)	6.21 (4.74)	.94 (.2.14)	1.12 (3.48) [.18 (2.90)]	−.14 (2.03)
Connectivity	.31 (.21)	.31 (.21)	.33 (.21)	.30 (.23)	.0 (.08)	.02 (.05) [.02 (.05)]	.0 (.07)
TROG-2 blocks	3.75 (4.43)	4.75 (4.71)	4.69 (4.78)	6.00 (5.18)	1.00 (2.45)	.94 (2.17) [−.06 (1.48)]	.46 (1.51)

3. Results

All participant testing sessions took place between October 2019 and March 2023, with an interruption of 18 months due to the COVID pandemic. As planned, the average time span between B1 and B2 for the Immediate group ($n = 17$) was 29 days (range = 22–42, SD = 3.96), while the Deferred group ($n = 16$) waited 57 days on average for their second baseline session (range = 50–63, SD = 2.80). All 33 participants completed the 12 therapy sessions over a period of 22–46 days (Immediate: $M = 30$ days, SD = 6.71 vs Deferred: $M = 27$ days, SD = 2.45). There was variability in the time span of intervention for some participants and thus, the intensity at which therapy was delivered. The reasons for the extended time spans were due to participant factors such as illness, and contextual factors such as transport strikes and public holidays. The mean gap for all participants between the end of therapy and the first outcome session (O1) was four days (SD = 3, range: 1–13). At follow-up, the two groups had similar time spans between O1 and O2: Immediate ($n = 14$): $M = 60.7$ days, SD = 7.27, range = 46–76; Deferred ($n = 13$): $M = 62.6$ days, SD = 11.84, range = 54–99. Again, both personal (illness, holidays) and contextual factors impacted on the availability of some participants at follow-up and the extended range of measurement points. Table 2 displays group data for the main outcome measures.

3.1. Connectivity outcomes

Higher connectivity scores indicate greater ability to combine words into grammatical trigrams. As a normative reference, connectivity measures obtained from a sample of 30 neurotypical adults narrating the "Dinner Party" cartoon demonstrated a mean connectivity score of .79 (SD = .05) (Zimmerer et al., 2017). Figure 3 shows Connectivity scores for both groups across the four assessment points. The figure suggests baseline stability in Connectivity scores, followed by modest improvement in the therapy phase, and then different profiles for the two groups in the maintenance phase with the Immediate group showing continuing change and the Deferred group falling back to baseline levels. One possibility is that Immediate group participants continue to improve during maintenance because they were less severe than the Deferred group.

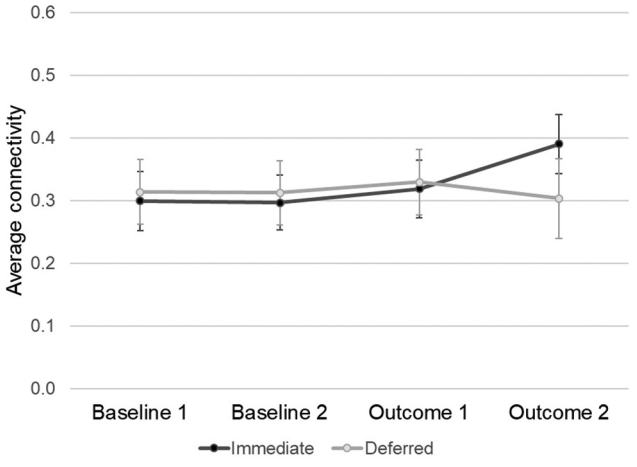

Figure 3. Mean connectivity over time by group (error bars = SEM).

3.1.1. Baseline stability

A two-way mixed ANOVA with Time (B1, B2) as the within-participant factor and Group (Immediate, Deferred) as a between-participant factor confirmed no significant differences across the baseline phase ($F(1,32) = .018$, $p = .894$, $\eta_p^2 = .001$ or between groups ($F(1, 32) = .235$, $p = .631$, $\eta_p^2 = .007$). There was no interaction effect ($F(1,32) = .007$, $p = .932$, $\eta_p^2 = .0$).

3.1.2. Usual care versus treatment

Next, we compared baseline performance (B1) to the timepoint 8 weeks later between groups, driven by our hypothesis that there may be a difference in speech connectivity between usual care and treatment groups following therapy. We exploited the treatment vs no-treatment RCT design and compared Connectivity scores for the Immediate group's therapy phase with scores across the Deferred group's extended 8-week no-treatment baseline phase (Figure 4). Connectivity was stable in the no-treatment period for the Deferred group, but there was a trend for improvement in the Immediate group. A 2×2

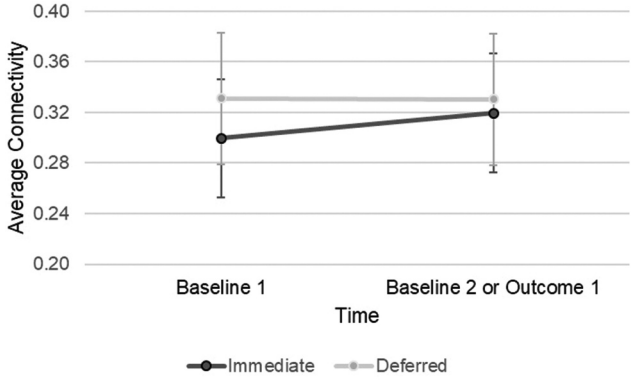

Figure 4. Mean connectivity over time: B1 versus B2 for Deferred group; B1 versus O1 for Immediate group (error bars = SEM).

mixed ANOVA with Time (B1, B2 for Deferred or O1 for Immediate) as a within-participant factor and Group as a between-participant factor was conducted, based on 17 participants in the Immediate group and 17 in the Deferred group. We note this is underpowered in relation to the recruitment target of 46 participants. Results show that there was no main effect of Time ($F(1,32) = .475$, $p = .496$, $\eta_p^2 = .015$), and no main effect of Group ($F(1,32) = .096$, $p = .758$, $\eta_p^2 = .003$) on Connectivity. There was also no interaction effect ($F(1,32) = .535$, $p = .470$, $\eta_p^2 = .016$). Thus, the results of this (underpowered) trial do not support our prediction that sentence therapy would improve speech connectivity more than usual care.

3.1.3. Post-hoc whole-group analysis of change over time

Given the underpowered nature of the study, we merged the two groups in a subsidiary LMM analysis to explore whether connectivity scores increased after therapy. This represents a post-hoc/unplanned analysis and data pooling is not an optimal strategy. However, given the exploration of a new therapy and the impact of the pandemic on the study, the analysis provided insight of the potential for further development. For more details on model comparison and diagnostics see Appendix 5. The results of the model comparison showed some evidence in favour of treating Time as a continuous versus categorical predictor (indicated by small differences in AIC of 3.33 (Model 1 v 2) and 3.32 (Model 3 v 4)), and Aphasia Severity as a categorical predictor (differences in AIC of 4.42 (Model 1 v 3) and 4.41 (Model 2 v 4); see Appendix 5). Model 1 was therefore selected as the winning model which parameterised Time and Aphasia Severity as continuous and categorical predictors respectively, whilst Model 2 was selected to provide contrasts between timepoints.

For Model 1, Time was a significant predictor of Connectivity with scores increasing over time, $B = 0.00891$, $t(29.89) = 2.12$, $p = 0.04$ (Table 3). Fixed effects showed model-estimated group differences in overall Connectivity according to Aphasia Severity. Specifically, individuals in the severe group had significantly lower estimated connectivity

Table 3. Linear mixed effects results for Connectivity (model 1).

Fixed Effects					
	Estimate	SE	t	df	p
Intercept	0.40	0.04	9.78	7.14	< 0.001***
Time (numerical)	0.01	0.00	2.12	29.89	0.04*
BNT-Moderate	0.06	0.05	1.19	14.76	0.25
BNT-Severe	−0.21	0.06	−3.32	13.81	0.005**
Random Effects					
				Variance	SD
Participant (Intercept)				0.03	0.16
Model fit					
				Marginal	Conditional
R^2				0.33	0.86

*$p < 0.05$, **$p < 0.01$ ***$p < 0.001$. BNT fixed effects are compared to the reference level BNT-Mild. Fixed effects calculated with robust standard errors using clubSandwich package in R. Model 1 equation: Connectivity predicted by Time (continuous) + BNT (categorical) + (1 | Participant).

Table 4. Planned Connectivity contrasts for timepoints (model 2).

Contrast	Estimate	Standard error	df	T ratio	p
B1 v B2	−0.0005	0.0070	215	−0.07	0.94
B2 v O1	0.0093	0.0071	215	1.33	0.19
O1 v O2	0.0030	0.0075	216	0.40	0.69
B v O1	0.0182	0.0122	215	1.49	0.14
B v O2	0.0241	0.0133	216	1.82	0.07
B v O	0.0212	0.0103	216	2.05	0.04*

*$p < 0.05$. B = baseline average; O = outcome average. Model 2 equation: Connectivity predicted by Time (categorical) + BNT (categorical) + (1 | Participant).

as compared to the mild group, B = −0.20806, t(13.81) = −3.32, $p = 0.005$, but there was no significant difference between the moderate and mild groups ($p = .253$).

Next, we performed planned contrasts for Model 2 (where Time was modelled as a categorical predictor) to examine differences in Connectivity scores between timepoints (Table 4). Overall, we found a significant increase in estimated Connectivity between baseline average and outcome average timepoints ($p = 0.04$). Individual contrasts were not statistically significant; however, the pattern of t-values is consistent with the expected direction of effects: stability over baseline and improvement in the therapy phase, with a small gain in the maintenance phase.

3.2. Sentence comprehension outcomes

TROG-2 scores are presented in Table 2 and Figure 5. Bishop (2003) reports normative TROG-2 scores as 18.48 blocks (SD = 1.69) in younger adults (17–64 years), and 18.07 blocks (SD = 1.86) in older adults (65–86 years). All participants showed marked spoken sentence comprehension impairment in comparison to these norms. The data suggest a learning effect on comprehension testing across the baseline phase. There was minimal change in the therapy period, followed by modest improvement in the maintenance phase.

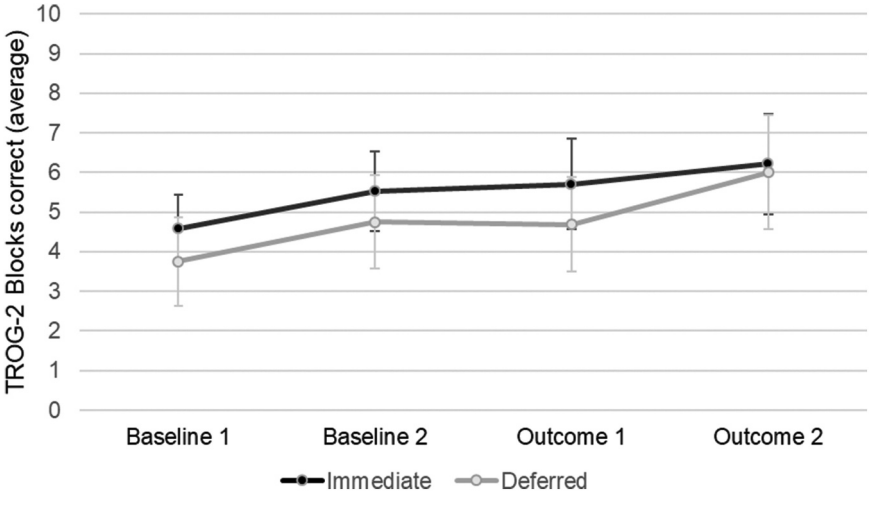

Figure 5. Mean TROG-2 Blocks correct over time, by group (error bars = SEM, max. score = 20).

3.2.1. Baseline stability

We first assessed baseline stability using a 2 × 2 mixed-factor ANOVA with Time (B1, B2) as the within-participant factor and Group (Immediate, Deferred) as a between-participant factor. There was no significant main effect of Group ($F(1, 32) = .050$, $p = .825$, $\eta_p^2 = .002$), but there was a significant main effect of Time ($F(1,32) = 5.426$, $p = .026$, $\eta_p^2 = .145$). Moreover, there was no significant interaction effect ($F(1,32) = .006$, $p = .932$, $\eta_p^2 = .0$). This result might reflect increasing familiarity with the test materials and researcher.

3.2.2. Usual care versus treatment

We again exploited the treatment vs no-treatment RCT design and compared TROG-2 scores across the Immediate group's therapy phase (B1 vs O1) with the Deferred group's extended 8-week no-treatment baseline phase (B1 vs B2; Figure 6). TROG-2 scores show a similar increase during the extended baseline period for the Deferred group and the treatment period in the Immediate group. We performed a 2 × 2 ANOVA with Time (B1, B2 for Deferred and B1, O1 for Immediate) as a within-participant factor and Group as a between-participant factor, based on 17 participants in the Immediate group and 17 in the Deferred group. Results indicated no main effect of Time ($F(1,32) = 3.784$, $p = .061$, $\eta_p^2 = .106$) or Group ($F(1,32) = .079$, $p = .780$, $\eta_p^2 = .002$). There was no interaction effect ($F(1,32) = .052$, $p = .820$, $\eta_p^2 = .002$).

3.2.3. Post-hoc whole-group analysis of change over time

Using LMMs, we compared TROG-2 scores over time. For more details on model comparison and diagnostics see Appendix 5. Model diagnostics indicated normality of residuals was violated due to two outlying data points. A secondary analysis was run with these data points excluded. This resolved non-normality and produced similar results, therefore we report the original analysis without exclusions.

Time was a significant predictor of TROG-2, with scores increasing over time, $B = 0.39$, $t(30.22) = 2.31$, $p = 0.03$ (Table 5). As with Connectivity, there were model-estimated group differences in TROG-2 scores according to Aphasia Severity, with the severe subgroup showing significantly lower estimated TROG-2 performance as compared to the Mild

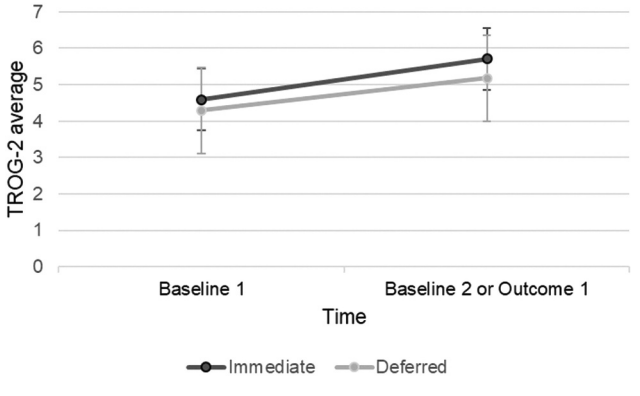

Figure 6. Mean TROG-2 Blocks over time: B1 versus B2 for Deferred group; B1 versus O1 for Immediate group (error bars = SEM).

Table 5. Linear mixed effects results for TROG-2 (model 5).

Fixed Effects					
	Estimate	SE	t	df	p
Intercept	8.22	1.63	5.03	7.27	0.001**
Time (numerical)	0.39	0.17	2.31	30.22	0.03*
BNT-Moderate	−1.44	2.10	−0.68	14.75	0.50
BNT-Severe	−6.37	1.74	−3.67	13.79	0.003**
Random Effects					
				Variance	SD
Participant (Intercept)				10.55	3.25
Model fit					
				Marginal	Conditional
R^2				0.39	0.87

*$p < 0.05$, **$p < 0.01$. BNT fixed effects are compared to the reference level BNT-Mild. Fixed effects calculated with robust standard errors using clubSandwich package in R. Model 5 equation: TROG-2 predicted by Time (continuous) + BNT (categorical) + (1 | Participant).

Table 6. Planned TROG-2 contrasts for timepoints (model 6).

Contrast	Estimate	Standard error	df	T ratio	p
B1 v B2	0.46	0.20	91.00	2.28	0.03*
B2 v O1	0.04	0.20	91.14	0.19	0.85
O1 v O2	0.13	0.22	91.42	0.57	0.57
B2 v O2	0.16	0.22	91.57	0.75	0.50
B2 v O	0.20	0.36	91.35	0.56	0.58

*$p < 0.05$; O = outcome average. Model 6 equation: TROG-2 predicted by Time (categorical) + BNT (categorical) + (1 | Participant).

subgroup, B = −6.37, t(13.79) = −3.67, $p = 0.003$, but there was no significant difference between the Moderate and Mild groups ($p = 0.50$).

Next, we performed contrasts on TROG-2 scores by timepoint (Table 6). There was a significant increase in estimated TROG-2 scores between B1 and B2 which was driving the significant effect of Time (p=0.03; see also section 3.2.1). To remove the influence of this effect from further contrasts, B2 was taken as the baseline reference. No other contrasts were significant (p>0.05)

3.3. Relationship between treatment outcome and baseline demographic and behavioural profiling variables

We conducted a correlational analysis to investigate the relationship between treatment outcome and baseline demographic and behavioural profiling variables such as age, years of education and naming. We used the whole-group average connectivity change score (O1-B1) as our treatment outcome variable. Connectivity-change scores did not show a significant linear relationship with any of the variables (Appendix 6).

3.4. QoL perceptions

QoL perceptions were collected at two timepoints: B2 and O2. A comparison across all participants, using a paired-samples t-test, showed no significant change in overall

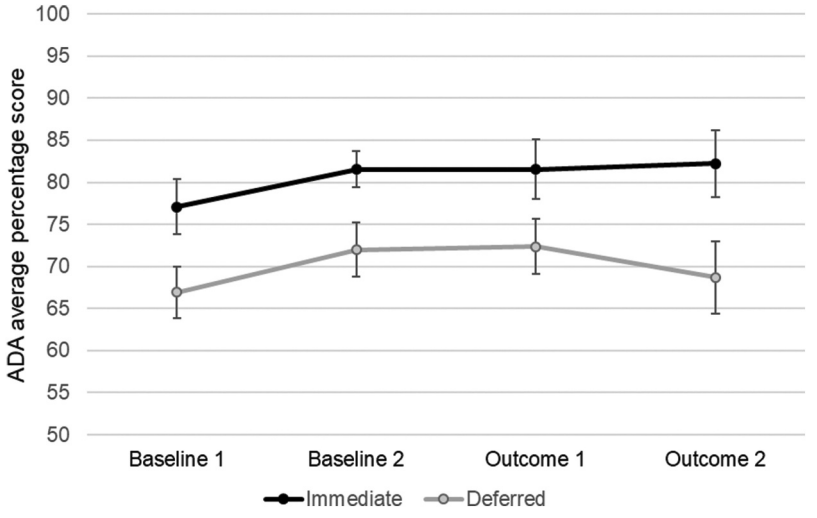

Figure 7. Mean Synonym Judgement percentage scores over time, by group (error bars = SEM).

SAQOL-39 ratings (B2: $M = 3.91$ (.60); O2: $M = 4.05$ (.56), $t(26) = -1.721$, $p = .097$, $d = .356$, small effect). However, scores on the SAQOL-39 communication sub-scale were significantly higher at maintenance ($M = 3.35$ (.88)) as compared to before therapy ($M = 2.9$ (.82); $t(26) = -3.35$, $p = .002$, $d = .65$, medium effect). This suggests that participants' self-perception of communication was enhanced.

3.5. Control measure

With regard to our untreated control measure (subset of the A.D.A. written synonym judgement task; Franklin et al., 1992), we predicted stable performance over time. First, we tested baseline stability, using a 2×2 mixed ANOVA with Time (B1, B2) as a within-subject factor and Group (Immediate, Deferred) as a between-subject factor. Results showed a significant main effect of Time ($F(1,31) = 9.247$, $p = .005$, $\eta_p^2 = .230$) and a significant main effect of Group ($F(1,31) = 6.407$, $p = .017$, $\eta_p^2 = .171$). Elevated scores at B2 (Immediate ($n = 17$): $M = 81.53$, $SD = 8.83$; Deferred ($n = 16$): $M = 72.0$, $SD = 13.02$) as compared to B1 (Immediate ($n = 17$): $M = 77.06$, $SD = 13.49$; Deferred ($n = 16$): $M = 66.94$, $SD = 12.26$) suggest increased familiarity with the test, while the main effect of Group indicated that scores were higher in the Immediate group consistent with their lower overall severity. There was no significant interaction ($F(1,31) = .036$, $p = .851$, $\eta_p^2 = .001$). Scores for both groups were stable between B2 and O1 (i.e., during the therapy phase; see Figure 7). A further 2×4 mixed ANOVA with Group (Immediate, Deferred) and Time (B1, B2, O1, O2) again revealed a group difference ($F(1,24) = 9.521$, $p = .005$, $\eta_p^2 = .284$), with higher scores in the Immediate group. There was also a main effect of Time ($F(3,72) = 3.664$, $p = .016$, $\eta_p^2 = .132$), but no significant interaction between factors ($F(3,72) = .516$, $p = .672$, $\eta_p^2 = .021$).

3.6. Acceptability

Intervention acceptability was high. The question "Overall, how helpful did you find this therapy?" (with highest possible rating of 5 = "very helpful") resulted in an average score of 4.77 (SD = 0.42) across all participants.

4. Discussion

This study sought to evaluate a usage-based sentence therapy for aphasia. The therapy programme included high-frequency sentences to target constructions used in everyday conversations. UTILISE is designed to enhance participants' auditory input processing, and in a subsequent speaking task, to loosen and lengthen semi-fixed phrases such as *[PERSON] like [THING]*. The study was disrupted by the COVID pandemic, which necessitated changes in protocol and had a negative impact on participant recruitment and retention. The study lacked statistical power in relation to the initial power calculation. The therapy was delivered in a clinic and at relatively low dose in relation to recommended levels (Bhogal et al., 2003; Brady et al., 2016). Participants received a maximum of 12 hours therapy, delivered in 12 sessions over a four-week period. However, despite operating at low dose, there was preliminary evidence for the potential of usage-based sentence therapy, constituting a sound basis for investigating the efficacy of the intervention at higher dosage.

Although underpowered, the sample size is larger than previous sentence therapy studies. Despite the intensive schedule of travel to a university clinic, the level of recruitment and attendance reflects the demand for therapy by people with chronic aphasia. Leaving aside issues linked to the pandemic, only two of 38 participants withdrew from the intensive regime. Furthermore, high acceptability ratings, as well as a significant improvement in self-perception on the communication subscale of the SAQOL-39, indicate a positive response to therapy directed at connected speech.

Contrary to our predictions, there was no significant difference in sentence production or comprehension scores when comparing a treated to a no-treatment (extended baseline) group using traditional ANOVA methods. However, when the full Connectivity data set was analysed using LMMs, we found a significant effect of Time on Connectivity, with participants performing significantly better at post-therapy timepoints (versus baseline timepoints). The pattern of change was consistent with trial phases suggestive of a treatment specific effect: baseline stability followed by a gain in the therapy phase and smaller gain in the maintenance phrase (however, it is important to note these changes did not reach statistical significance). The TROG-2 analysis revealed a different pattern; participants displayed significant improvement over time, but this was driven by improvements over baseline, suggestive of a familiarity effect, followed by relatively smaller and non-significant increases over therapy and maintenance, suggesting changes relating to time rather than therapy. Both LMMs (Connectivity and TROG-2) included Aphasia Severity as a factor, showing that participants with severe aphasia tended to have lower estimated Connectivity and TROG-2 scores.

The significant LMM results contrast the non-significant ANOVA results. This could be partly explained by the addition of Aphasia Severity in the LMM analyses which accounted for different levels of language ability. As noted previously, this method also has several

advantages, notably that we were able to make use of a richer data set by including multiple responses per participant and partial data which maximised power. Our results are in line with a previous comparison of these techniques in the aphasia population which suggests LMMs may be more sensitive than ANOVAs in identifying subtle differences in outcomes (Mohapatra & Dash, 2023).

Notably there was a small increase in Connectivity in the maintenance phase, whereas previous studies tend to show a loss of therapy gains in this phase. This increase might be due to continued improvement in Connectivity or, alternatively, participants may have been tired at the immediate outcome point after the four-week intensive regime of travel and sessions. The O1 session took place, on average, at four days after the end of therapy. Typically, outcomes are measured close to the cessation of therapy to avoid fall-off of treatment effects. However, in more intensive regimes there may be a need to balance proximity to treatment with allowing participants to rest when considering the timing of post-treatment outcome measures.

The main outcome measures employed in the study were distal measures, in that neither the narrative nor TROG-2 tasks directly targeted the high-frequency constructions that featured in the therapy. It is difficult to probe spontaneous production of treated items using traditional measures such as picture description, particularly given the focus on high-frequency, yet semantically light/low-imageability verbs. Therefore, inherently, the outcome measures reported here involve generalisation to untreated constructions, as well as transfer to other tasks. We also included a study-specific story completion test in which we probed both the treated structures and within-level generalisation to length/frequency-matched untreated structures at baseline and O1 timepoints. The results for this more proximal outcome measure are reported in Bruns et al. (in prep.).

In this study, we used the FLAT software to profile connected speech production. The automated analysis delivers a number of count, ratio and frequency variables. We selected trigram connectivity as a primary outcome measure based upon small-scale pilot studies and a view that it reflects the capacity to combine words into grammatical constructions and equates roughly to a clinical impression of fluency. Other FLAT variables may be useful in intervention studies, such as word and construction frequency, type-token ratios, and ratios of content-function words. FLAT has been used to characterise production in both development and a range of acquired language impairments (Sederias et al., 2024; Zimmerer et al., 2020), but its use is at an early stage in outcome research (Bruns et al., 2021). The automated analysis has particular advantages for intervention research in that it is fast, and outputs are masked, avoiding researcher bias in evaluations. However, data preparation is time consuming both in orthographic transcription of audio samples and in tagging text. Although both these steps could be subject to rater bias, we avoided such issues by masking independent transcribers/taggers to phase of data collection, and producing a manual to define procedure, training and inter-rater reliability checks, to maximise data fidelity.

In terms of limitations, the study was disrupted by the COVID pandemic which impacted recruitment and retention of participants, and necessitated a change of protocol after study onset. We recruited participants via a university research register and participation in the study placed high demands on study volunteers and their families in terms of intensive travel and attendance. Both these factors may result in a biased sample and recruitment of participants who are highly motivated to

contribute to research and seek further therapy. All the outcome measures were administered by a researcher who was not masked to phase of the assessment. Although we incorporated procedures for data fidelity checks into the study design, ideally outcomes would be assessed by a researcher masked to phase; funding limitations precluded this option.

At an impairment-level, therapy research and practice has been dominated by single-word naming interventions, with comprehension and sentence-level impairments receiving less attention. This study offers an innovative approach to clinical intervention for sentence processing impairments in aphasia, based on application of a new framework to therapy. Given its focus on high-frequency constructions that are used in everyday interactions, this approach offers the potential of meaningful functional change in language abilities. This preliminary study has revealed some encouraging trends regarding gains in spontaneous speech production. The findings remain tentative, however, this approach merits future investigation and elaboration to explore its effectiveness and utility as a clinical intervention for aphasia.

Notes

1. Naming therapy search terms: ((naming) OR (anomia) OR (lexical) OR (vocabulary)) AND ((treatment) OR (therapy) OR (intervention)) AND ((aphasia) OR (agrammatism)); sentence therapy search terms: ((sentences) OR (sentences) OR (syntax) OR (grammar)) AND ((treatment) OR (therapy) OR (intervention)) AND ((aphasia) OR (agrammatism)).
2. This resulted in inclusion of one additional participant in the Deferred group, who subsequently dropped out before therapy began.

Acknowledgements

We thank Michael Coleman for software programming; Ewa Dąbrowksa for her advice on selecting stimuli; Richard Jardine for his time in making audio recordings; Vitor Zimmerer for access and support in use of the FLAT software; Helena Illing, India Bamrah and Mi Xuan Chua for conducting the literature searches on word versus sentence therapy, and Tae Horsfield for the TROG-2 fidelity check.

Disclosure statement

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

Funding

The study was supported by the Stroke Association under Grant [TSA SA PG 18\100054]. The views expressed are those of the author(s) and not necessarily those of the funder.

ORCID

Claudia Bruns http://orcid.org/0000-0002-0835-5485
Michael Dean http://orcid.org/0000-0003-2690-641X

Jane Warren http://orcid.org/0000-0002-9858-124X
Victoria Fleming http://orcid.org/0000-0003-2582-4914
Amir-Homayoun Javadi http://orcid.org/0000-0003-0569-6441
Rosemary Varley http://orcid.org/0000-0002-1278-0601

Author contributions

RV and CB wrote the manuscript; KD, FR, MD, AJH, VF and JW contributed to the revision of the manuscript. All authors read and agreed the final version of the manuscript.

Availability of data and materials

LMM R code and dataset available from https://osf.io/j9udn/. Anonymised data (subject to participant consent to data sharing) are available from RV in response to reasonable request.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bello-Lepe, S., Mahmood, S., Varley, R., & Zimmerer, V. (2024). Speech pauses in speakers with and without aphasia: A usage-based approach. *Cortex, A Journal Devoted to the Study of the Nervous System and Behavior*, *178*, 287–298. https://doi.org/10.1016/j.cortex.2024.06.012

Berthier, M. L. (2005). Poststroke aphasia: Epidemiology, pathophysiology and treatment. *Drugs and Aging*, *22*(2), 163–182. https://doi.org/10.2165/00002512-200522020-00006

Berthier, M. L., & Pulvermüller, F. (2011). Neuroscience insights improve neurorehabilitation of poststroke aphasia. *Nature Reviews: Neurology*, *7*(2), 86–97. https://doi.org/10.1038/nrneurol.2010.201

Bhogal, S. K., Teasell, R., & Speechley, M. (2003). Intensity of aphasia therapy, impact on recovery. *Stroke*, *34*(4), 987–992. https://doi.org/10.1161/01.STR.0000062343.64383.D0

Bishop, D. V. M. (2003). *Test for reception of grammar, version 2 (TROG-2)*. Pearson.

BNC. (2007). *BNC, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium*.

Bolognini, N., Russo, C., & Edwards, D. J. (2016). The sensory side of post-stroke motor rehabilitation. *Restorative Neurology and Neuroscience*, *34*(4), 571–586. https://doi.org/10.3233/RNN-150606

Brady, M., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*, *Issue* (6. Art. No.: CD000425. https://doi.org/10.1002/14651858.CD000425.pub4

Bruns, C. (in prep). *Measuring outcomes of sentence production therapy: Probing constructions with a story completion test*.

Bruns, C., Beeke, S., Zimmerer, V. C., Bruce, C., & Varley, R. A. (2021). Training flexibility in fixed expressions in non-fluent aphasia-a case series report. *International Journal of Language & Communication Disorders*, *56*, 1009–1025. https://doi.org/10.1111/1460-6984.12652

Bruns, C., Varley, R., Zimmerer, V. C., Carragher, M., Brekelmans, G., & Beeke, S. (2019). "I don't know": A usage-based approach to familiar collocations in non-fluent aphasia. *Aphasiology*, *2*(33), 140–162. https://doi.org/10.1080/02687038.2018.1535692

Brysbaert, M., & Ellis, A. W. (2016). Aphasia and age of acquisition: Are early-learned words more resilient? *Aphasiology*, *30*(11), 1240–1263. https://doi.org/10.1080/02687038.2015.1106439

Burgess, P. W., & Shallice, T. (1997). *The Hayling and Brixton tests: Two tests of dysexecutive syndrome*. Thames Valley Test Company.

Chan, A.-W., Tetzlaff, J. M., Gøtzsche, P. C., Altman, D. G., Mann, H., Berlin, J. A. & Krleža-Jeric, K. (2013). SPIRSPIRIt, 2013 explanation and elaboration: Guidance for protocols of clinical trials. *British Medical Journal (clinical Research*, *346*, 1–42. https://doi.org/10.1136/bmj.e7586

Chomsky, N. (1981). *Lectures on government and binding: The pisa lectures*. Walter de Gruyter.
Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074
Conroy, P., Sotiropoulou Drosopoulou, C., Humphreys, G. F., Halai, A. D., & Lambon Ralph, M. A. (2018). Time for a quick word? The striking benefits of training speed and accuracy of word retrieval in post-stroke aphasia. *Brain : A Journal of Neurology*, *141*(6), 1815–1827. https://doi.org/10.1093/brain/awy087
Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001
Dąbrowska, E. (2014). Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics*, *25*(4), 617–653. https://doi.org/10.1515/cog-2014-0057
Difrancesco, S., Pulvermüller, F., & Mohr, B. (2012). Intensive language-action therapy (ILAT): The methods. *Aphasiology*, *26*(11), 1317–1351. https://doi.org/10.1080/02687038.2012.705815
Edmonds, L. A. (2016). A review of verb network strengthening treatment: Theory, methods, results, and clinical implications. *Topics in Language Disorders*, *36*(2), 123–135. https://doi.org/10.1097/TLD.0000000000000088
Edmonds, L. A., Nadeau, S. E., & Kiran, S. (2009). Effect of verb network strengthening treatment (VNeST) on lexical retrieval of content words in sentences in persons with aphasia. *Aphasiology*, *23*(3), 402–424. https://doi.org/10.1080/02687030802291339
Faul, E., Erdfelder, F., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149
Fillingham, J. K., Sage, K., & Lambon Ralph, M. A. (2006). The treatment of anomia using errorless learning. *Neuropsychological Rehabilitation*, *16*(2), 129–154. https://doi.org/10.1080/09602010443000254
Fletcher, M., & Birt, D. (1983). *Storylines: Picture sequences for language practice*. Longman.
Franklin, S., Turner, J., & Ellis, A. (1992). *The A.D.A. Auditory comprehension battery*. University of York.
Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496. https://doi.org/10.1353/lan.0.0035
Gahl, S., Menn, L., Ramsberger, G., Jurafsky, D. S., Elder, E., Rewega, M., & Holland Audrey, L. (2003). Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences. *Brain & Cognition*, *53*(2), 223–228. https://doi.org/10.1016/S0278-2626(03)00114-3
Goldberg, A. E. (1995). *A construction grammar approach to argument structure*. The University of Chicago Press.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001
Goodglass, H., Gleason, J. B., Ackerman Bernholtz, N., & Hyde, M. R. (1972). Some linguistic structures in the speech of a Broca's aphasic. *Cortex, A Journal Devoted to the Study of the Nervous System and Behavior*, *8*(2), 191–212. https://doi.org/10.1016/S0010-9452(72)80018-2
Hartsuiker, R. J., & Kolk, H. H. J. (1998). Syntactic facilitation in agrammatic sentence production. *Brain and Language*, *62*(2), 221–254. https://doi.org/10.1006/brln.1997.1905
Helm-Estabrooks, N. A., Nicholas, M., & Morgan, A. (1989). *Melodic intonation therapy program*. Pro-Ed.
Hickin, J., Cruice, M., & Dipper, L. (2019). A systematically conducted scoping review of the evidence and fidelity of treatments for verb deficits in aphasia: Verb-in-isolation treatments. *American Journal of Speech Language Pathology*, *29*(1S), 530–559. https://doi.org/10.1044/2019_ajslp-cac48-18-0234
Hickin, J., Cruice, M., & Dipper, L. (2022). A systematically conducted scoping review of the evidence and fidelity of treatments for verb and sentence deficits in aphasia: Sentence treatments. *American Journal of Speech Language Pathology*, *31*(1), 431–462. https://doi.org/10.1044/2021_AJSLP-21-00120
Hilari, K., Byng, S., Lamping, D. L., & Smith, S. C. (2003). Stroke and aphasia quality of life scale-39 (SAQOL-39): Evaluation of acceptability, reliability, and validity. *Stroke*, *34*(8), 1944–1950. https://doi.org/10.1161/01.STR.0000081987.46660.ED

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D. & Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal (Online)*, *348*(March), 1–12. https://doi.org/10.1136/bmj.g1687

IBM Corp. Released. (2023). *IBM SPSS statistics for windows, version 29.0.2.0*.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test* (2nd ed.). Lippincott Williams & Wilkins.

Kay, J., Coltheart, M., & Lesser, R. (1992). *Psycholinguistic assessments of language processing in aphasia*. Lawrence Erlbaum Associates.

Kaye, R. C., & Cherney, L. R. (2016). Script templates: A practical approach to script training in aphasia. *Topics in Language Disorders*, *36*(2), 136–153. https://doi.org/10.1097/tld.0000000000000086

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology*, *25*(4), 463–492. https://doi.org/10.1080/02643290701674851

Lam, J. M. C., & Wodchis, W. P. (2010). The relationship of 60 disease diagnoses and 15 conditions to preference-based health-related quality of life in Ontario hospital-based long-term care residents. *Medical Care*, *48*(4), 380–387. https://doi.org/10.1097/MLR.0b013e3181ca2647

Lee, J., & Man, G. (2017). Language recovery in aphasia following implicit structural priming training: A case study. *Aphasiology*, *31*(12), 1441–1458. https://doi.org/10.1080/02687038.2017.1306638

Leff, A. P., Nightingale, S., Gooding, B., Rutter, J., Craven, N., Peart, M., Dunstan, A., Sherman, A., Paget, A., Duncan, M., Davidson, J., Kumar, N., Farrington-Douglas, C., Julien, C., & Crinion, J. T. (2021). Clinical effectiveness of the queen square intensive comprehensive aphasia service for patients with poststroke aphasia. *Stroke*, *52*(10), e594–e598. https://doi.org/10.1161/STROKEAHA.120.033837

Marshall, J., Devane, N., Edmonds, L., Talbot, R., Wilson, S., Woolf, C., & Zwart, N. (2018). Delivering word retrieval therapies for people with aphasia in a virtual communication environment. *Aphasiology*, *32*(9), 1054–1074. https://doi.org/10.1080/02687038.2018.1488237

Mohapatra, B., & Dash, T. (2023). Linear mixed-model analysis better captures subcomponents of attention in a small sample size of persons with aphasia. *American Journal of speech-Language Pathology*, *32*(2), 748–761. https://doi.org/10.1044/2022_AJSLP-22-00119

Nickels, L., & Howard, D. (1995). Aphasic naming: What matters? *Neuropsychologia*, *33*(10), 1281–1303. https://doi.org/10.1016/0028-3932(95)00102-9

Palmer, R., Dimairo, M., Cooper, C., Enderby, P., Brady, M., Bowen, A. & Chater, T. (2019). Self-managed, computerised speech and language therapy for patients with chronic aphasia post-stroke compared with usual care or attention control (big CACTUS): A multicentre, single-blinded, randomised controlled trial. *Lancet Neurology*, *18*(9), 821–833. https://doi.org/10.1016/S1474-4422(19)30192-9

Pustejovsky, J. E. (2023). *ClubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. R package version 0.5.8*. https://CRAN.R-project.org/package=clubSandwich

R Core Team. (2025). *R: A language and environment for statistical computing*. URL: R Foundation for Statistical Computing. https://www.R-project.org/

Ross, K. B., & Wertz, R. T. (2003). Quality of life with and without aphasia. *Aphasiology*, *17*(4), 355–364. https://doi.org/10.1080/02687030244000716

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010, statement: Updated guidelines for reporting parallel group randomised trials. *British Medical Journal (Online)*, *340*(7748), 698–702. https://doi.org/10.1016/j.ijsu.2010.09.006

Sederias, I., Krakovitch, A., Stojanovik, V., & Zimmerer, V. C. (2024). Overuse of familiar phrases by individuals with Williams syndrome masks differences in language processing. *Journal of Child Language*, *27*, 1–15. https://doi.org/10.1017/S0305000924000436. Advance online publication.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *37*(3), 776–784. https://doi.org/10.1037/a0022531

Standard Occupational Classification. (2020). *Office for national statistics*. https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020/soc2020volume1structureanddescriptionsofunitgroups

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

Thompson, C. K., Choy, J. J., Holland, A., & Cole, R. (2010). Sentactics®: Computer-automated treatment of underlying forms. *Aphasiology*, *24*(10), 1242–1266. https://doi.org/10.1080/02687030903474255

Thompson, C. K., Shapiro, L., Kiran, S., & Sobecks, J. (2003). The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia: The complexity account of treatment efficacy (CATE). *Journal of Speech, Language, and Hearing Research*, *42*, 690–707. https://doi.org/10.1044/1092-4388(2003/047)

Upton, E., Doogan, C., Fleming, V., Leyton, P. Q., Barbera, D., Zeidman, P., Hope, T., Latham, W., Coley-Fisher, H., Price, C., Crinion, J., & Leff, A. (2024). Efficacy of a gamified digital therapy for speech production in people with chronic aphasia (iTalkbetter): Behavioural and imaging outcomes of a phase II item-randomised clinical trial. *EClinicalMedicine*, *70*, 102483. https://doi.org/10.1016/j.eclinm.2024.102483

Varley, R. (2011). Rethinking aphasia therapy: A neuroscience perspective. *International Journal of speech-Language Pathology*, *13*(1), 11–20. https://doi.org/10.3109/17549507.2010.497561

Varley, R., Bruns, C., Warren, J., Dąbrowska, E., & Javadi, A.-H. (2020, March 18). Computer therapy combined with non-invasive brain stimulation for sentence processing difficulties in post-stroke aphasia: A randomised control trial (the UTILISE study). *Open Science Framework Preprints*. https://doi.org/10.31219/osf.io/fduqh

Varley, R., Bruns, C., Warren, J., Dąbrowska, E., Rodgers, F., & Javadi, A.-H. (2021). Addendum to the article "computer therapy combined with non-invasive brain stimulation for sentence processing difficulties in post-stroke aphasia: A randomised control trial (the UTILISE study). https://osf.io/qvpjm

Varley, R., Cowell, P. E., Dyson, L., Inglis, L., Roper, A., & Whiteside, S. P. (2016). Self-administered computer therapy for apraxia of speech: Two-period randomized control trial with crossover. *Stroke*, *47*(3), 822–828. https://doi.org/10.1161/STROKEAHA.115.011939

Veale, J. F. (2014). Edinburgh Handedness Inventory - short form: A revised version based on confirmatory factor analysis. *Laterality: Asymmetries of Body, Brain and Cognition*, *19*(2), 164–177. https://doi.org/10.1080/1357650X.2013.783045

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. https://doi.org/10.1080/02687038.2015.1027169

Wechsler, D. (2011). *Wechsler abbreviated scale of intelligence - second edition (WASI-II)*. Psychological Corporation.

Whiteside, S. P., Inglis, A. L., Dyson, L., Roper, A., Harbottle, A., Ryder, J. & Varley, R. A. (2012). Error reduction therapy in reducing struggle and grope behaviours in apraxia of speech. *Neuropsychological Rehabilitation*, *22*(2), 267–294. https://doi.org/10.1080/09602011.2011.639614

Whiteside, S. P., & Varley, R. A. (1998). A reconceptualisation of apraxia of speech: A synthesis of evidence. *Cortex, A Journal Devoted to the Study of the Nervous System and Behavior*, *34*(2), 221–231. https://doi.org/10.1016/s0010-9452(08)70749-4. PMID: 9606587.

Wilshire, C. E., Lukkien, C. C., & Burmester, B. R. (2014). The sentence production test for aphasia. *Aphasiology*, *28*(6), 658–691. https://doi.org/10.1080/02687038.2014.893555

Wray, A. (2002). Patterns of Formulaicity in aphasic language. In A. Wray (Ed.), *Formulaic language and the lexicon* (pp. 217–246). Cambridge University Press.

Zimmerer, V. C., Coleman, M., Hinzen, W., & Varley, R. A. (2017, September). Reliance on common word combinations correlates with degree of syntactic impairment in aphasia. *Stem-Spraak-En Taalpathologie*. *Supplement*, *2: Abstracts 18th International Science of Aphasia Conference* (Vol. *22*. pp. 163–164, *Geneva*.

Zimmerer, V. C., Hardy, C. J. D., Eastman, J., Dutta, S., Varnet, L., Bond, R. L., Russell, L., Rohrer, J. D., Warren, J. D., & Varley, R. A. (2020). Automated analysis of language production in aphasia and right-hemisphere damage: Frequency and collocation strength. *Cortex, A Journal Devoted to the Study of the Nervous System and Behavior*, *133*, 103–119. https://doi.org/10.1016/j.cortex.2020.08.027

Zimmerer, V. C., Newman, L., Thomson, R., Coleman, M., & Varley, R. (2018). Automated analysis of language production in aphasia and right hemisphere damage: Frequency and collocation strength. *Aphasiology*, *32*(11), 1267–1283. https://doi.org/10.1080/02687038.2018.1497138

Zimmerer, V. C., Wibrow, M., & Varley, R. A. (2016). Formulaic language in people with probable Alzheimer's disease: A frequency-based approach. *Journal of Alzheimer's Disease*, *53*(3), 1145–1160. https://doi.org/10.3233/JAD-160099