Factive and counterfactive interpretation of embedded clauses in aphasia and its relationship with lexical, syntactic and general cognitive capacities

Zimmerer, V.C.[a], Varley, R.A.[a], Deamer, F.[b], Hinzen, W.[c,d]

[a]Department of Language and Cognition, University College London

[b]Department of Philosophy, Durham University

[c]Institute for Research and Advanced Studies (ICREA)

[d]Department of Translation and Language Sciences, Universitat Pompeu Fabra

Corresponding author:

Vitor Zimmerer

Department of Language and Cognition

Division of Psychology and Language Sciences

UCL

v.zimmerer@ucl.ac.uk

Chandler House

2 Wakefield Street

London WC1N 1PF

Accepted for *Journal of Neurolinguistics* on 16/08/18

Highlights:

- We tested aphasic comprehension of factives, non-factives and counterfactives.
- Impairment was greater in trials which required counterfactive interpretation.
- Performance in all trials correlated with degree of language impairment.
- Performance in counterfactive trials also correlated with non-verbal reasoning.

Abstract:

In factive clausal embedding ([*He knows* [*that it is warm outside*]]), the embedded clause is presupposed to be true. In non-factive embedding ([*He thinks* [*that it is warm outside*]]) there is no presupposition, and in counterfactive embedding ([*It only seems* [*that it is warm outside*]]) the embedded clause is presupposed to be false. These constructions have been investigated as a window into the complexity of language and thought, and there are disputes as to the relative contributions of lexical, syntactic or non-verbal resources in their interpretation. We designed a sentence-picture matching task to test comprehension of these constructions in a group of aphasic participants and in neurotypical controls. In particular, we tested the capacity to reach a factive or counterfactive interpretation. In factive interpretation trials, participants with aphasia performed nearly as well as controls, while in counterfactive interpretation trials they performed significantly worse. Accuracy in factive and counterfactive interpretation trials correlated with other syntactic and lexical measures. Only performances on counterfactive trials correlated with non-verbal reasoning measures. Exploratory regression models suggest that verbal and non-verbal scores were separate factors. Results indicate that a disruption of counterfactive interpretation in aphasia is linked to reduction of syntactic and/or conceptual-propositional capacities.

Keywords: aphasia; grammar; factives; counterfactives; propositions; reasoning

1. Introduction

You and your friend get ready for a day trip to the coast. You see your friend pack light clothing and sunscreen. “I know that it is sunny over there,” she says. You check the weather forecast. It is all clouds and rain. “No,” you say. “You just think that it is sunny.”

Lexically and syntactically, *She knows that it is sunny outside* and *She thinks that it is sunny outside* appear similar. Both sentences have the same argument structure, a tail-embedded copular clause, a high-frequency mental verb in the matrix clause and the same inflections. However, their propositional structures differ substantially (Kiparsky & Kiparsky, 1970; Sheehan & Hinzen, 2011).

Without further context, the clause *It is sunny outside* is assumed to be true. This default interpretation is not affected in a factive sentence like *She knows that it is sunny* outside. For this reason, *She knows that it is sunny outside, but it is not sunny* is contradictory. The interpretation of the embedded clause as true is overridden in non-factives constructions. In a sentence like *She thinks that it is sunny outside*, the embedded clause can be false, and in counterfactives (a subset of non-factives), such as *You just think that it is sunny outside*, the listener even expects it to be false. These types of clausal embedding appear across constructions (e.g., interrogatives: *Do you know there's juice in the fridge?* vs. *Do you think there's juice in the fridge?*). They communicate an individual's mental state and the reliability of information. They involve complex processing at several cognitive levels and have been investigated in order to determine how these levels might interact.

In this report we examine the capacity of people with aphasia and non-brain damaged (NBD) controls to generate the correct factive and counterfactive interpretation of embedded clauses based on the verb phrase in the matrix clause. Research in aphasia has informed theories of language processing (e.g., Gahl & Menn, 2016; Grodzinsky, 2000) and the relationship between language and thought (Apperly, Samson, Carroll, Hussain, & Humphreys, 2006; Baldo, Paulraj, Curran, & Dronkers, 2015; Blank, Balewski, Mahowald, & Fedorenko, 2016; Varley, 2014; Varley, Klessinger, Romanowski, & Siegal, 2005; Varley & Siegal, 2000). We investigated the nature of aphasic comprehension in trials that require either factive and counterfactive interpretations; the degree to which it is impaired, and how possible impairment relates to other aspects of cognition.

Because comprehension of these constructions has not been investigated in aphasia, we looked at explanations from the child development literature as well as related findings from aphasia to establish a theoretical framework. Comprehension of factives, non- and counterfactives in embedding has been studied extensively in child language (e.g., see Dudley, Orita, Hacquard, & Lidz, 2015, for a review) in order to learn how maturation of different cognitive mechanisms contributes to eventual understanding of these constructions. The age at which full comprehension is achieved is not clear. Some studies suggest that successful differentiation starts at year 3 or 4 (e.g., Dudley et al., 2015; Johnson & Maratsos, 1977; Lewis, Hacquard, & Lidz, 2012), but Léger (2007) indicates that full insight into factivity is not complete until age 11. Results depend much on the methods employed. Dudley et al. (2015) criticize tests of factivity comprehension for often relying on metalinguistic reasoning, such as judging the appropriateness of sentences given a verbally presented context (e.g., Falmagne, Gonsalves, & Bennett-Lau, 1994; Harris, 1975) or adding additional cognitive demands by simultaneously assigning multiple mental states to different characters (Léger, 2007).

Dudley et al. (2015), in their review of the literature, list four explanations for this phenomenon: (1) Conceptual demands, including Theory of Mind (ToM), i.e. the requirement to attribute to an individual thoughts that may be different from one's own. (2) Syntactic demands, as the embedded clause must be integrated within the matrix clause. (3) Interpretation of the pragmatic context within which the construction is produced (for example, *think* is often used parenthetically instead of referring to a [possibly] false belief); (4) Lexical knowledge of the role the matrix verb plays in assigning (non-)factivity. These hypotheses concern different aspects of language processing and do not have to be mutually exclusive.

Hypotheses with much explanatory power in child development may be less powerful for adult aphasia, and vice versa, given that individuals with aphasia experience impairment to matured language networks. Reviewing the aphasia literature, syntactic and lexical accounts appear more likely than those which concern social reasoning and context. Clause integration is often disrupted in aphasia, with extensive evidence of difficulties in processing subject and object relatives (Caramazza & Zurif, 1976; Friedmann & Gvion, 2003; Swinney & Zurif, 1995). Lexical-semantic processing, of both nouns and verbs, is also impaired (Druks, 2002; Berndt, Mitchum, Haendiges, & Sandson, 1997). By contrast, there is evidence for retained social reasoning in aphasia. A series of studies with severely aphasic individuals (chance performance in sentence comprehension tasks and almost no connected language output) have shown good performances on tests designed to test non-verbal ToM and communication (Varley, Siegal, & Want, 2001; Varley & Siegal, 2000; Willems, Benn, Hagoort, Toni, & Varley, 2011; Zimmerer & Varley, 2010). Apperly et al. (2006) reported good performance of PH, a man with syntactic impairment who showed retained capacity in non-verbal first- and second-order ToM tasks. However, he also succeeded in a verbal test in which he had to answer questions with non-factive constructions such as *Where does Jeremy think the bag is?,* and counterfactual questions like *What if the waitress had not noticed the bag?*. It could therefore be argued that PH had at least some access to linguistic resources, which he could have used in the non-verbal tasks.

Bánreti, Hoffman and Vincze (2016) had participants verbally report mental states represented in pictured situations. They found that participants with aphasia successfully communicated mental states. However, instead of producing utterances with embedded clauses, they tended to report them in first person direct speech, as if quoting the person in the picture. People with a diagnosis of Alzheimer's disease on the other hand had more clause embedding in their output, but more often failed to convey relevant ToM content. In a discussion of previous studies and their own data the authors argue for a double dissociation between clause embedding and ToM processing. However, there have been no investigations of the question to what degree aphasic social cognition involves

full representations of others' mental states, as opposed to more perceptual or action-oriented cognition (Butterfill & Apperly, 2013; Rubio-Fernandez & Geurts, 2013).

Beyond ToM, cognitive demands may also include the general ability to maintain complex propositional representations. As mentioned earlier, a listener would be biased to interpret an utterance like *It is sunny outside* as true. As an embedded clause in a non-factive or counterfactive context, this bias competes with the correct interpretation. In addition to common lexical and syntactic processing demands across factives and non- or counterfactives, the latter demand inhibition and manipulation of propositional content. Duman, Altınok, & Maviş (2016) suggest that an impairment of a "general cognitive capacity" can occur in aphasia, with particular disruption of executive function, resulting in impaired comprehension of counterfactual if-clauses in Turkish compared to comprehension of factual if-clauses. This proposal is based on Duman et al.'s claim that their factual and counterfactual stimuli are equivalent with regards to morphological and syntactic complexity, which would rule out linguistic impairment as the reason for this dissociation. General cognitive impairment has been associated with aphasia (Baldo et al., 2015; Peristeri & Tsimpli, 2013), though there are reported cases of people with very severe aphasia and strong non-verbal reasoning skills (Varley et al., 2005; Zimmerer, Cowell, & Varley, 2014).

We approached the current investigation with three questions:

(Q1) Is the ability to generate factive and counterfactive interpretations of embedded clauses impaired in participants with aphasia?

(Q2) Is aphasic comprehension of embedded clauses poorer when a counterfactive interpretation needs to be reached?

(Q3) How does comprehension of these clauses relate to other verbal and non-verbal capacities both in participants with aphasia and NBD controls?

We designed a sentence-picture matching (SPM) task to test comprehension of factivity, which allows us to place our results within the wider context of sentence comprehension research in aphasia. We had two trial types: factive interpretation trials and counterfactive interpretation trials. Factive interpretation trials used factive constructions as stimuli, and the matching picture showed the embedded clause to be true. Counterfactive interpretation trials used non- or counterfactive constructions, and the matching picture showed the embedded clause to be false. For example, for the trial sentence *The man thinks that it is warm outside* the correct picture showed that the weather was cold.

We selected four matrix constructions that are frequent in everyday use. We list them in ranked order, starting with the most factive. The first two were used in factive interpretation trials, the last two in counterfactive interpretation trials.

1. *Know* construction (NP *knows that* S): *Know* is considered a factive (e.g., Dudley et al., 2015).
2. *It is clear* construction (*It is clear to* NP *that* S): This construction is typically interpreted as factive. However, it does not withstand a negation test which is seen as a test of full factivity (Kiparsky & Kiparsky, 1970): *It isn't clear to the man that it is sunny outside* strongly suggests, but does not entail, that it is sunny outside.
3. *Think* construction (NP *thinks that* S): *Think* is considered a paradigmatically non-factive verb. The complement may or may not be true. In our experiment, matching pictures required the presupposition that it is false.
4. *It only seems* construction (*It only seems to* NP *that* S): This construction is counterfactive as the complement clause is presupposed to be false.

Our data were collected as part of a larger project ("Language and Mental Health") which addresses a range of questions about language in cognitive disorders, of which comprehension of factive, non-factive and counterfactive embedding is only one. Test protocols included extensive language and cognitive testing (see Appendix A for our protocol). To address Q3 about the contribution of verbal and non-verbal capacities in understanding our sentences, we correlated factive and counterfactive trial performance with a selection of other cognitive measures (however, we did not test ToM or pragmatic capacity). Our choice of measures can be linked to three of the cognitive requirements discussed above: syntactic, lexical, and general cognitive demands. Tests were selected before conducting the analyses reported in this article.

When testing conceptual cognitive capacities in aphasic participants, it is important to choose non-verbal tests. If tests contain too much verbal material aphasic participants are likely to fail because of their language impairment, regardless of intellectual capacity. We selected three tests of non-verbal capacities: The three picture version of Pyramids and Palm Trees (PPT; Howard & Patterson, 1992) as a test of non-linguistic semantic ability, the Wechsler Abbreviated Scale of Intelligence (Wechsler, 2011) Matrices subtest, which assesses non-verbal reasoning, and the Brixton Spatial Anticipation Test (Brixton; Burgess & Shallice, 1997) to assess executive function. Our protocol also included the Ravens Coloured Progressive Matrices (RCPM, Raven, Raven, & Court, 2004), which is similar to WASI II Matrices. We chose WASI Matrices over the RCPM to avoid redundancy and because RCPM data set also showed a strong ceiling effect.

As a test of syntactic processing we chose the Test for Reception of Grammar version 2 (TROG-2; Bishop, 2003). The TROG-2 is a SPM task testing a range of different constructions including canonical and non-canonical sentences, relative clauses, different types of negation and comparatives. Our protocol also contained syntactic assessments from the Comprehensive Aphasia Test (CAT; Howard, Swinburn, & Porter, 2004), which are also conducted via SPM. We chose the TROG-2 over CAT sentences to avoid redundancy and because our sample included some participants with mild aphasia which resulted in scores at ceiling in the easier CAT subtest.

To assess lexical capacities in production we included the Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 2001), and in comprehension, the spoken word-picture matching test from the Comprehensive Aphasia Test (CAT; Howard, Swinburn, & Porter, 2004). We also tested verbal working memory using the digit span recognition subtest of the Psycholinguistic Assessment of Language Processing in Aphasia (PALPA13; Kay, Lesser, & Coltheart, 1997).

Our hypotheses are based on the findings reviewed above: Language impairment is the defining feature of aphasia, non-verbal cognitive capacities can be affected, and counterfactive interpretations being possibly harder because of greater reliance on clausal integration.

To address Q1, we hypothesized that

(H1) participants with aphasia will have lower SPM accuracy than controls.

To address Q2, we hypothesized that

(H2a) accuracy of participants with aphasia will be higher in factive than in counterfactive trials,

(H2b) differences to controls will be larger in counterfactive trials

To address Q3, we hypothesized that people with better performances in standardized tests would achieve higher SPM accuracy. We formulated hypotheses to test the effects of (H3) lexical capacity, measured by tests of lexical production (BNT) and comprehension (CAT spoken words), (H4) syntactic capacity, measured by the test of sentence comprehension (TROG-2), (H5) verbal working memory, measured using digit span recognition (PALPA13) and (H6) non-verbal capacities, measured using tests of non-verbal reasoning and association (PPT, WASI-II Matrices, Brixton). We tested H3-6 separately for each group. Note however that we tested aphasic participants, but not NBD controls, on (CAT) spoken word comprehension, with consequences for significance thresholds (see Results).

Finally, we tested age and education effects within control groups, hypothesizing (H7) a negative correlation between age and SPM performance and (H8) a positive correlation between years of formal education and SPM performance.

Accepted for *Journal of Neurolinguistics* on 16/08/18

The study was granted ethical approval from an institutional ethics committee UCL (LC/2013/05). All participants gave informed consent to taking part in the study.

2. Experiment

2.1. Participants

We recruited 21 participants with aphasia and 30 NBD controls. Given current criticisms of the classical aphasia model and linked syndromes as neurologically and behaviorally inconsistent (Berndt & Caramazza, 1999; Caramazza, Capitani, Rey, & Berndt, 2001; Tremblay & Dick, 2016), we recruited a heterogeneous group (Tables 1 and 2). Instead of using subgroup analyses, we examined the effect of impairment at various levels by correlating test scores with SPM performance.

Aphasic participants were recruited via convenience sampling through UCL's communication clinic and contacts at UK Connect. Controls were recruited via London chapters of the University of the Third Age. Participants had normal or corrected-to-normal vision and no reported hearing impairment. We had to exclude one aphasic participant because she had difficulties understanding a number of tasks and showed signs of stress. The remaining 20 participants with aphasia had a mean age of 63.7 (SD = 10.72), and 16 were male. The control group had a mean age of 70.5 (SD = 7.01), and 9 were male. Age differences were significant, $t(29.79) = 2.409$, $p = .022$, $d = .78$ and we therefore included age as a covariate in analyses. On the basis of the participants' employment and their highest academic attainment, we estimated years of formal education. For participants with aphasia, the mean was 13.86 (SD = .47); for controls, it was 14.27 (SD = .36). The difference was not significant, $t(49) = -.7$, $p = .486$. Participants were classified as fluent or non-fluent by an experienced clinician (Varley) on the basis of recordings of spontaneous speech. Table 1 shows an overview of all aphasic participants.

Testing was conducted across three sessions. Two controls did not attend the final test session and therefore scores for TROG-2, WASI-II Matrices and Brixton were missing. For six controls there was an experimenter error in applying the BNT stopping criterion. As it was applied too early, these participants were invited back for retesting. One participant could not be retested and we excluded her BNT data. Controls were not tested on word comprehension and one can expect ceiling performances in this group.

We calculated ANCOVAs with age as a covariate to compare groups across standardized tests (see Table 2 for descriptive and full inferential values). The Brixton score usually is a count of errors, meaning that higher scores indicate worse performance. To facilitate comparisons to other tests for

which higher scores indicate better performance, we inverted Brixton scores so that they represent correct responses. WASI-II Matrices scores are raw scores (number of matrices solved). As one would expect, aphasic performance was significantly lower ($p < .001$) in all comparisons of verbal behavior (BNT; TROG-2; PALPA 13), with large effect sizes ($r > .6$). The only non-verbal test on which groups significantly differed was the Brixton, with the control group showing a slightly better performance than participants with aphasia ($p = .048$). There was a trend towards higher scores by the control group on WASI-II Matrices ($p = .053$). We found age to have a significant effect on WASI-II scores, but not on other measures.

Table 1. Overview of background data of aphasic participants. In all cases, aphasia resulted from a stroke. Years PO = Years post-onset

ID	Age	Sex	Profession	Years of education	Years PO	Aphasia description
1	81	M	Teacher	12	2	Mild fluent
2	50	M	Operation manager	14	3	Non-fluent
3	53	M	Engineer	15	2	Mild fluent
4	77	M	Pharmacologist	18	5	Non-fluent
5	50	F	Accountant	14	8	Non-fluent
6	46	F	Accountant	14	2	Non-fluent
7	56	M	Driver	10	5	Mild non-fluent
8	58	F	Designer	16	9	Non-fluent
9	54	M	Operation manager	12	2	Non-fluent
10	71	M	Fire researcher	12	8	Non-fluent
11	56	M	Upholsterer	12	3	Non-fluent
12	72	M	Head of school	18	25	Fluent
13	66	M	Engineer	12	5	Fluent
14	83	M	Teacher	15	12	Fluent
15	73	M	Designer	14	6	Non-fluent
16	60	M	Decorator	16	4	Non-fluent
17	70	F	Biologist	14	9	Non-fluent
18	67	M	Accountant	14	5	Non-fluent
19	67	M	Electrician	10	3	Non-fluent
20	63	M	Computer consultant	15	9	Non-fluent

Table 2. Standardized test scores for aphasic and control groups. Higher scores indicate better performance.

Test	Aphasia (SD)	Controls (SD)	Comparison (age effect)
Object naming (BNT; max = 60)	40.8 (15.4)	57.1 (2.6)	$F(1,46) = 29.03$, $p < .001$, $r = .62$ ($p = .695$, $r = .07$)
Spoken word comprehension (CAT; max = 30)	27.3 (2.2)	-	-
Spoken sentence comprehension (TROG-2; max = 20)	12 (5)	19 (.9)	$F(1,45) = 51.802$, $p < .001$, $r = .72$ ($p = .334$, $r = .14$)
Digit span recognition (PALPA 13; max = 7)	4.9 (1.2)	6.4 (.6)	$F(1,47) = 34.411$, $p < .001$, $r = .65$ ($p = .348$, $r = .14$)
Non-verbal semantics (PPT; max = 52)	50.2 (2.2)	50.5 (1.5)	$F(1,47) = .132$, $p = .718$, $r = .05$ ($p = .48$, $r = .1$)
Non-verbal reasoning (WASI-II Matrices; max = 30)	17.1 (4.5)	18.3 (2.8)	$F(1,45) = 3.958$, $p = .053$, $r = .27$ ($p = .039$, $r = .29$)
Non-verbal executive function (Brixton; max = 55)	35.5 (5.5)	38 (5.6)	$F(1,44) = 4.151$, $p = .048$, $r = .28$ ($p = .96$, $r = .24$)

BNT: Boston Naming Test

CAT: Comprehensive Aphasia Test

TROG-2: Test of Reception of Grammar

PALPA 13: Psycholinguistic Assessments of Language Processing in Aphasia

PPT: Pyramids and Palm Trees

WASI-II: Wechsler Abbreviated Scale of Intelligence

Brixton = Brixton Spatial Anticipation Test (number of correct trials)

max = maximum score attainable

Accepted for *Journal of Neurolinguistics* on 16/08/18

2.2. Materials and procedure

Materials

Our materials are available for download (www.cognitionandgrammar.net/s/Factivity-SPM.zip). The stimulus set included 35 arrays of three black and white drawings. Each array was presented vertically on a sheet of A4 paper. For each array a stimulus sentence was matched to one of the three pictures. We used four different matrix constructions (see introduction). In line with our aim to keep sentences lexically and syntactically simple, we chose copular constructions for the complement clauses. See Table 3 for a full list of sentences.

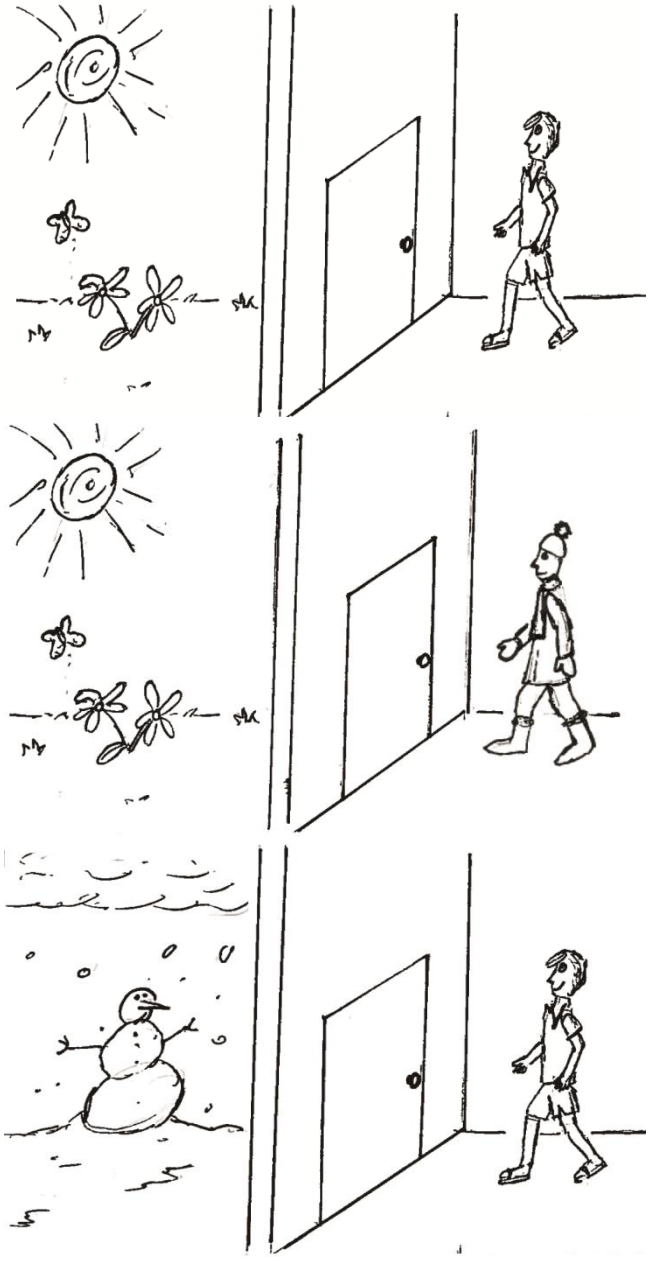
Pictures were based on ten different scenarios (or event types). Five scenarios were used for *know* and *think* trials. Five scenarios were used for *it is clear* and *it only seems* trials. Each scenario contained a) an experiencer and b) a situation. The experiencer was referred to as grammatical subject of the matrix clause, the situation was referred to in the embedded clause, e.g., *The man knows that it is warm outside* or *The man thinks that it is warm outside.* Pictures were drawn using a "doll house" perspective. The experiencer was always on the right side of the picture. In *know* and *think* trials, the experiencer was separated from the scenario by a wall and depicted in a way that his or her mental representation of the event could be inferred. Figure 1 shows one example: In this scenario, the situation is the weather outside being sunny and warm, or snowy and cold, and the experiencer is a man who wears either light clothing (t-shirt, shorts, flip flops) or heavy clothing (coat, long trousers, boots, scarf, gloves, beanie hat). In addition, facial expressions of the experiencer contributed to making mental representations interpretable. The man with light clothing smiles, the man with heavy clothing does not. We expected participants to infer that the former assumes that it is warm outside, while the latter assumes that it is cold.

Matching sentences to the correct reference picture in *know* trials (e.g. *The man knows that it is warm outside*; Figure 1) required a factive interpretation. The sentences only fit pictures in which the situation and the representation of the experiencer matched the proposition in the embedded clause (e.g., it is warm outside and the man wears light clothing). Arrays contained the target and two distractors: a) experiencer match + situation mismatch and b) experiencer mismatch + situation match.

In *think* trials (e.g., *The man thinks that it is cold outside*; Figure 1) the matching picture corresponded with the counterfactive interpretation, and the situation did not match the embedded clause (e.g., the man wears light clothing while it is cold outside). By the nature of this trial, both

distractors were always experiencer mismatches. They were characterized as: a) experiencer mismatch + situation match and b) experiencer mismatch + situation mismatch.

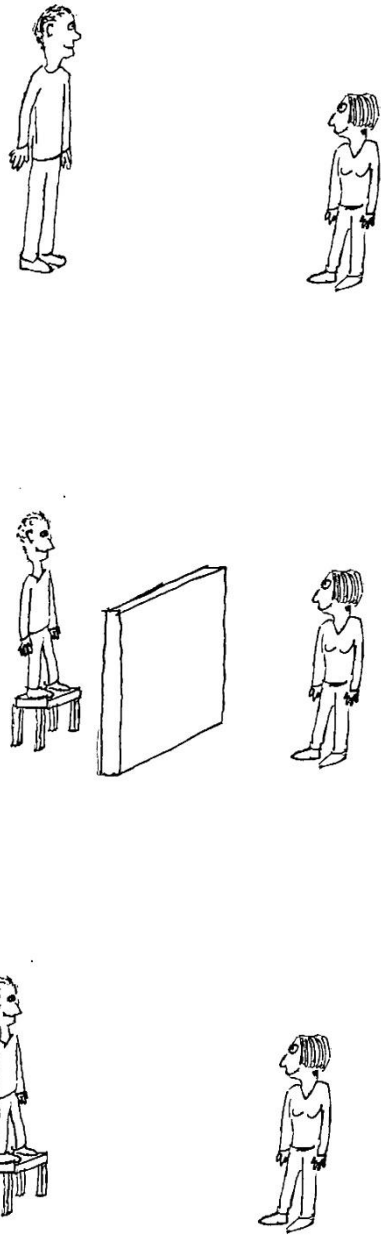
Figure 1. Example for a picture array for *know* and *think* trials. We used this particular combination of pictures for two sentences: *The man knows that it is warm outside* (top picture correct; factive interpretation) and *The man thinks that it is cold outside* (center picture correct; counterfactive interpretation). The bottom picture served as a distractor in the counterfactive trial (embedded clause correct) and, in another combination, as the target picture for *The man thinks that it is cold outside*. Target picture position was balanced across trials.



In *it is clear* and *it only seems* trials, it was necessary to interpret the perspective of the experiencer, which was either false (in pictures matching *it only seems* sentences) or direct and clear (*it is clear* trials). *It is clear* trials (e.g., *It is clear to the woman that the man is* tall; Figure 2) required factive interpretation. The sentences fit pictures in which the situation matched the embedded proposition and the experiencer could see the situation. Arrays contained the target and two distractors: a) experiencer perspective match + situation mismatch and b) experiencer perspective mismatch + situation mismatch.

Sentences in *it only seems* trials (e.g., *It only seems to the woman that the man is tall*; Figure 2) were counterfactive, meaning that selection of the matching picture required counterfactive interpretation. Arrays contained the target picture and two distractors: a) experiencer perspective mismatch + situation mismatch and b) experiencer state mismatch + situation match. During the design and piloting phase it proved difficult to clearly visualize some *it only seems* sentences. For instance, *It only seems to the woman that the computer was fixed* was easily depictable (see online supplement), but *It only seems to the woman that the computer was broken* was not. As a result, we included five sentences for these trials, while the other trials contained ten sentences each.

Figure 2. Example for a picture array for *it is clear* and *it only appears* trials. We used this particular combination of pictures for three sentences: *It is clear to the woman that the man is tall* (top picture correct), *It is clear to the woman that the man is small* (bottom picture correct) and *It only seems to the woman that the man is tall* (center picture correct). Target picture position was balanced across trials.



Procedure

Participants were tested in a quiet room in the communication clinic of the University College London, with the exception of one aphasic participant who was tested at home due to his limited

mobility. The experimenter showed the participant one array at a time. With each array, the stimulus sentence was read aloud by the experimenter. The experimenter aimed to read all stimuli with the same speed and intensity. Participants had to point at the picture which matched the spoken sentence. Participants could ask for a single repetition. Immediate self-corrections were allowed. The order of pictures was randomized in a way that each position was the target for approximately the same number of trials (top = 12 trials, center = 11 trials, bottom = 12 trials). Trials were randomized so that each scenario occurred once before scenarios were repeated. The same randomized order was used for all participants.

2.3. Results

2.3.1. Item analysis and exclusion

The first step in the analysis was to examine if any items elicited high levels of error. In the control group, mean accuracy for each item was 92.2%, SD = 9.5. Table 4 lists all trials including the percentage of correct responses in the control group. We identified two items for which control performance was 2 SDs below the item mean (threshold: 73.03%).

Table 3. All sentence stimuli used in the SPM task. “Trial no.” indicates trial order. Control accuracy = Percentage of correct responses for each test item. Items printed in bold present outliers (2 SDs below control mean).

Construction	Sentence	Trial no.	Control accuracy
Know **construction**	The mother knows that the child is naughty.	1	73.3%
	The mother knows that the child is nice.	8	100.0%
	The man knows that it is cold outside.	11	96.7%
	The man knows that it is warm outside.	16	96.7%
	The man knows that the dog is harmless.	23	93.3%
	The woman knows that dinner is ready.	24	96.7%
	The woman knows that dinner is not ready.	28	96.7%
	The man knows that the bathroom is clean.	31	100.0%
	The man knows that the dog is dangerous.	33	83.3%
	The man knows that the bathroom is dirty.	34	100.0%
Think **construction**	The man thinks that the dog is dangerous.	3	90.0%
	The man thinks that the bathroom is clean.	4	86.7%
	The man thinks that the dog is harmless.	10	93.3%
	The woman thinks that dinner is not ready.	**12**	**70.0%**
	The woman thinks that dinner is ready.	14	93.3%
	The man thinks that the bathroom is dirty.	17	100.0%
	The mother thinks that the child is naughty.	20	100.0%
	The man thinks that it is cold outside.	22	96.7%
	The man thinks that it is warm outside.	25	93.3%
	The mother thinks that the child is nice.	30	100.0%
It is clear **construction**	It is clear to the woman that the computer is fixed.	2	100.0%
	It is clear to the man that the dog is small.	5	100.0%
	It is clear to the woman that the stall is free.	6	100.0%
	It is clear to the woman that the man is tall.	9	96.7%
	It is clear to the woman that the man is small.	13	96.7%
	It is clear to the woman that the stall is occupied.	**18**	**66.7%**
	It is clear to the man that the pool is safe.	19	93.3%
	It is clear to the woman that the computer is broken.	21	93.3%
	It is clear to the man that the pool is dangerous.	29	80.0%
	It is clear to the man that the dog is big.	35	96.7%
It only seems **construction**	It only seems to the man that the pool is safe.	7	73.3%
	It only seems to the man that the dog is big.	15	90.0%
	It only seems to the woman that the man is tall.	26	100.0%
	It only seems to the woman that the stall is occupied.	27	100.0%
	It only seems to the woman that the computer is fixed.	32	80.0%

It is not clear why performance in some trials was poorer. Possible factors are the specific words used, the design of the visual stimuli or its implementation. Since these factors do not concern the ability to process factivity per se, we excluded the two outlier trials from analysis. For comparison, Table 5 contains means for each group both with and without outliers. Means summarize

performance across construction types, as well as trial types categorized by factive/counterfactive interpretation and the total average across all trials. Removing outliers changed accuracy only marginally, with no significant differences. We continued with outlier trials removed.

Table 4. Average accuracy (and SD) in aphasic and control groups. Accuracy averaged by construction (left side) and trial type (right side).

	know	***think***	***it is clear***	***it only seems***	**Factive interpretation trials**	**Counterfactive interpretation trials**	**Total**
Controls	93.7% (9.6)	92.3% (11.4)	92.3% (9.4)	88.7% (17.2)	93% (7.3)	91.1% (11.3)	92.2% (7.2)
Outlier items excluded	93.7% (9.6)	94.8% (9.5)	95.2% (7.5)	88.7% (17.2)	94.4% (6.8)	92.6% (10.2)	93.6% (6.5)
Aphasia	89.4% (24)	63.8% (34.7)	86.3% (23.1)	71.3% (28)	87.8% (22.4)	66.3% (31.1)	75.4% (17.9)
Outlier items excluded	89.4% (24)	65.3% (35.4)	90.3% (22.9)	71.3% (28)	89.8% (22.6)	67.4% (31.4)	76.9% (17.5)

Table 5. Overview of standardized test scores, factive and counterfactive SPM scores for aphasic participants. For SPM scores, outlier items were excluded (see Item Analysis section in 2.2).

ID	BNT	CAT (words)	TROG-2	PALPA 13 (digit span)	PPT	Brixton	WASI-II Matrices	Factive SPM trials	Counterfactive SPM trials
1	48	30	16	6	51	36	16	100%	93%
2	53	29	15	6	51	37	19	79%	86%
3	19	26	3	4	50	41	22	95%	79%
4	57	30	17	5	51	34	19	100%	71%
5	23	26	7	5	43	37	9	89%	57%
6	25	24	11	6	51	25	20	89%	50%
7	57	28	19	7	52	49	22	100%	100%
8	54	26	15	6	50	35	19	100%	79%
9	52	26	18	4	52	33	22	100%	100%
10	53	30	6	4	51	29	18	95%	79%
11	45	30	14	4	52	42	15	100%	86%
12	8	24	4	4	51	32	22	79%	86%
13	41	24	6	5	51	40	14	89%	14%
14	14	30	5	3	49	37	7	79%	7%
15	28	28	14	5	47	27	16	84%	71%
16	40	25	14	3	47	39	18	58%	21%
17	51	27	12	7	50	30	10	84%	29%
18	50	26	17	5	52	38	21	100%	79%
19	47	28	11	4	52	35	15	95%	64%
20	50	28	15	5	50	33	16	96%	20%

BNT: Boston Naming Test

CAT: Comprehensive Aphasia Test

TROG-2: Test of Reception of Grammar

PALPA 13: Psycholinguistic Assessments of Language Processing in Aphasia

PPT: Pyramids and Palm Trees

WASI-II: Wechsler Abbreviated Scale of Intelligence

Brixton: Brixton Spatial Anticipation Test (number of correct trials)

2.3.2. A priori hypotheses

Our a priori hypotheses were directional. For this reason, we report one-tailed p values.

(H1) Participants with aphasia will have lower accuracy than controls in our factivity SPM task.

Because of non-parametric distributions, we used a Rank Analysis of Covariance (Quade, 1967) with age as a covariate. In this analysis, the ranked outcome variable is residualized over the ranked covariate. Residuals are then compared between groups using independent t-tests. Overall, participants with aphasia performed worse than controls, $t(48) = 3.831$, $p < .001$, $d = 1.22$.

(H2a) accuracy of participants with aphasia will be higher in factive than in counterfactive trials,

(H2b) differences to controls will be larger in counterfactive trials

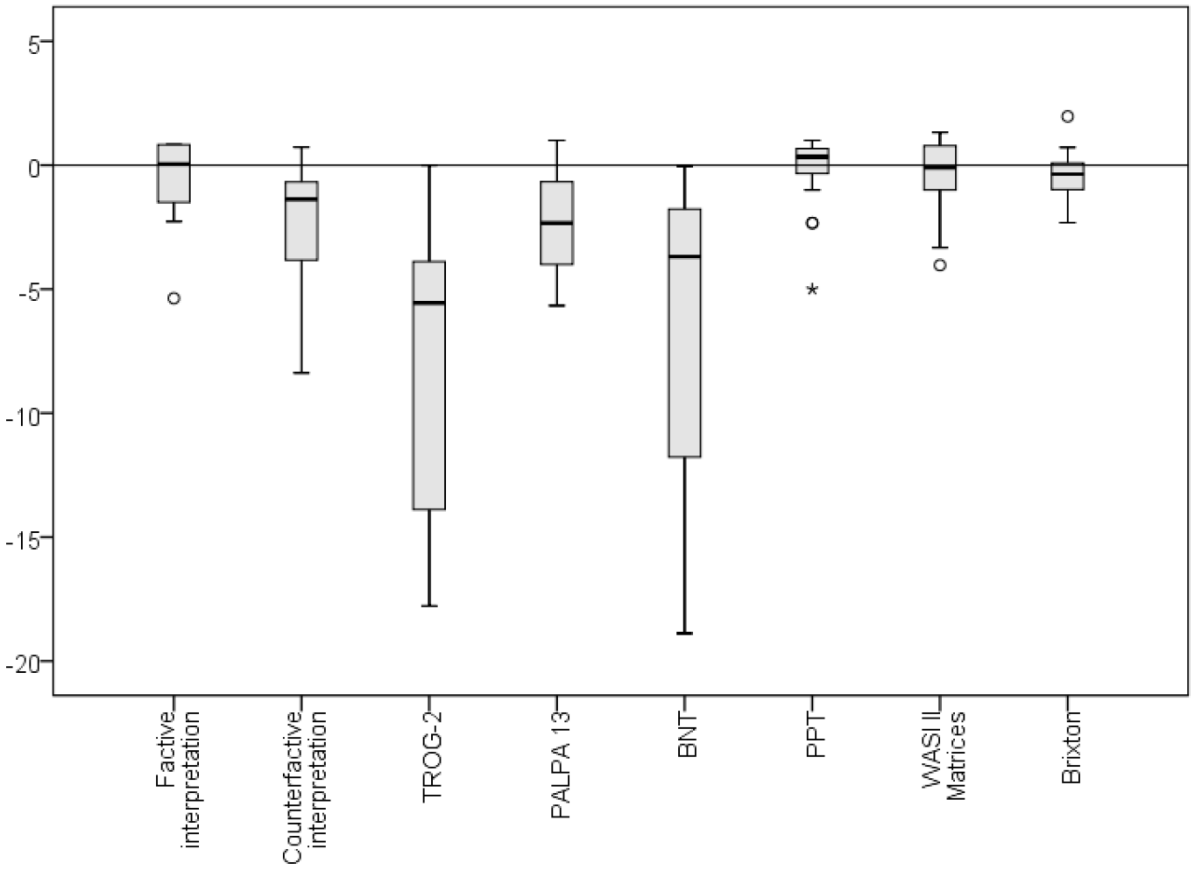
Quade's rank analysis of covariance showed no significant differences between groups in *know* trials, $t(48) = 1.202$, $p = .1$, $d = .4$, and *it is clear* trials, $t(48) = 1.438$, $p = .076$, $d = .46$. Differences were significant in *think* trials, $t(48) = 5.712$, $p < .001$, $d = 1.82$, and *it only seems* trials, $t(48) = 2.75$, $p = .004$, $d = .84$. These effects withstood Bonferroni correction (adjusted significance threshold $p = .0125$). When grouped into factive and counterfactive interpretation trials, differences between groups in factive interpretation trials were not significant, $t(48) = 1.442$, $p = .078$, $d = .47$, while differences in counterfactive interpretation trials were, $t(48) = 4.706$, $p < .001$, $d = 1.39$.

Within NBD controls, Friedman's ANOVA showed no significant differences between performance on different constructions, $\chi^2(3) = 2.082$, $p = .556$. When grouped into factive and counterfactive interpretation trials, Wilcoxon tests also showed no significant difference, $Z = .774$, $p = .439$, $r = .01$.

Within the aphasic group, there were significant differences between trial types, $\chi^2(3) = 29.566$, $p < .001$. Wilcoxon tests revealed that performance in *know* trials was better than *think* trials, $Z = 1.75$, $p < .001$, $r = .28$, and *it only seems* trials, $Z = 1.1$, $p = .007$, $r = .17$. Performance in *it is clear* trials was also stronger than *think* trials, $Z = 1.6$, $p < .001$, $r = .25$, and *it only seems* trials, $Z = .95$, $p = .02$, $r = .15$. Except for the latter, all differences were significant at a Bonferroni-corrected threshold ($p = .008$). When grouped into factive and counterfactive interpretation trials, Wilcoxon tests indicated that comprehension of factive embedding was better than counterfactive interpretation trials, $Z = 3.595$, $p < .001$, $r = .57$. The data illustrate poorer aphasic performance in counterfactive interpretation trials, with no performance differences between constructions within each trial type.

To put aphasic SPM performance into context, we converted aphasic scores from all tests into z-scores using the control group's mean and SD. Average z-scores for the aphasic group were -.53(SD = 1.63) for factive interpretation, -2.5 (SD = 2.8) for counterfactive interpretation, -7.78 (SD = 5.75) for the TROG, -2.5 (SD = 1.49) for the PALPA 13, -6.29 (SD = 5.92) for the BNT, -.23 (SD = 1.49) for PPT, -.43 (SD = 1.61) for WASI Matrices and -.45 (SD = 1.02) for the Brixton. Figure 3 visualizes the aphasic dataset in relation to performance of NBD controls.

Figure 3. Aphasic test scores and distributions in relation to control group data. Individual scores were z-transformed on the basis of control group means and SDs. The y-axis tracks performance in relation to the control group mean (origin), measured in SDs. Negative scores indicate weaker performance. Dots denote outliers from the control group (more than 1.5 x IQR), while asterisks denote far outliers (3 x IQR). All between-group differences were significant, with the exceptions of factive interpretation, PPT and WASI II Matrices (see also Table 2).



TROG-2: Test of Reception of Grammar

PALPA 13: Psycholinguistic Assessments of Language Processing in Aphasia

BNT: Boston Naming Test

CAT: Comprehensive Aphasia Test

PPT: Pyramids and Palm Trees;

WASI-II: Wechsler Abbreviated Scale of Intelligence

Brixton: Brixton Spatial Anticipation Test.

(H3-H6) Relationship between aphasic comprehension of factive and counterfactive interpretation and other test scores

We ran a series of correlations within the aphasic group to determine the relationship between SPM performance and other test variables. Because of the directional hypotheses we computed one-tailed correlations, and report all correlations with $p < .05$. Because of significant differences in performance between factive and counterfactive interpretation trials, we correlated standardized test scores against these trial types separately. However, factive and counterfactive accuracy were also strongly correlated, $\tau = .51$, $p = .002$ and conceptually related. In such a case, standard Bonferroni corrections are inappropriately strict (Perneger, 1998). We apply a solution suggested by Sankoh, Huque and Dubey (1997; for further discussion, see McKenzie, 2012), where p values are adjusted not using the number of comparisons *n*, but $n^{1-r(.k)}$, where r(.k) is the average correlation between the outcomes (in our data (.k) = $2^{1 - .51} = 1.4$). Distributions for our outcome variables were non-parametric according to Shapiro-Wilk tests. We therefore used Kendall's τ for correlations.

(H3) Lexical capacity (BNT, CAT spoken words)

We adjusted the significance threshold for comparisons with the two outcomes (see above) and two standardized tests which served as independent variables (adjusted threshold $p = .05 / (2 \times 1.4) = .018$). There were positive correlations between BNT scores and accuracy in factive trials, $\tau = .433$, $p = .007$, and accuracy in counterfactive trials, $\tau = .29$, $p = .041$. The latter correlation was not significant after Bonferroni correction. There were no significant correlations between CAT spoken word comprehension and accuracy in factive trials, $\tau = .154$, $p = .202$, or accuracy in counterfactive trials, $\tau = .113$, $p = .262$.

(H4) Syntactic capacity (TROG-2)

We adjusted the significance threshold for the two outcomes ($p = .05 / 1.4 = .036$). There were positive correlations between TROG-2 scores and accuracy in factive trials, $\tau = .428$, $p = .008$, and accuracy in counterfactive trials, $\tau = .429$, $p = .006$.

(H5) Verbal working memory (PALPA 13)

We adjusted the significance threshold for the two outcomes ($p = .05 / 1.4 = .036$). There were no significant correlations between PALPA 13 and accuracy in factive trials, $\tau = .191$, $p = .155$, or accuracy in counterfactive trials, $\tau = .14$, $p = .218$.

(H6) Non-verbal reasoning (PPT, WASI-II Matrices, Brixton)

We adjusted the significance threshold for the two outcome variables and three standardized tests which served as independent variables (adjusted threshold p = .05 / (3 x 1.4) = .012). PPT correlated with accuracy in factive trials, τ = .434, p = .01, and accuracy in counterfactive trials, τ = .395, p = .014. The latter correlation was marginally above the adjusted threshold. The correlation between performance on WASI-II Matrices with accuracy in factive trials was not significant τ = .226, p = .103. Its correlation with accuracy in counterfactive trials was significant τ = .452, p = .004. There were no significant correlations between Brixton scores and accuracy in factive trials, τ = .101, p = .282, or accuracy in counterfactive trials, τ = .05, p = .384.

In summary, SPM accuracy correlated significantly with performance in various standardized tests. Participants with higher scores on the PPT, BNT and TROG-2 were more accurate in factive interpretation trials. With regard to counterfactive interpretation trials, participants with higher scores on the WASI-II Matrices, TROG-2 and PPT were more accurate (although the latter correlation was barely above the adjusted significance threshold). There was an additional, weaker trend suggesting a relationship between BNT performance and accuracy in counterfactive trials.

(H3-H6) Relationship between NBD comprehension of factive and counterfactive interpretation and other test scores.

We conducted similar correlational analyses for NBD controls. The correlation between performance in factive and counterfactive trials in the control group was weak and not significant, τ = .142, p = .227. On the basis of this correlation the adjusted number for correction for multiple comparisons is 1.81. For the full and unadjusted correlations, see Appendix B.

(H3) Lexical capacity (BNT)

We adjusted the significance threshold for the two outcome variables (p = .05 / 1.81 = .028). There was no significant correlation between BNT scores and accuracy in factive trials, τ = .046, p = .381. The correlation between BNT scores and accuracy in counterfactive trials was significant, τ = .335, p = .014.

(H4) Syntactic capacity (TROG-2)

We adjusted the significance threshold for the two outcome variables (p = .05 / 1.81 = .028). There were no significant correlations between TROG-2 scores and accuracy in factive, τ = .058, p = .364, or counterfactive trials, τ = .127, p = .226.

(H5) Verbal working memory (PALPA 13)

We adjusted the significance threshold for the two outcome variables (p = .05 / 1.81 = .028). There were no significant correlations between PALPA 13 and accuracy in factive τ = .189, p = .129, or counterfactive trials, τ = .233, p = .082.

(H6) Non-verbal capacities (PPT, WASI-II Matrices, Brixton)

We adjusted the significance threshold for the two outcome variables and three standardized tests which served as independent variables (adjusted threshold p = .05 / (3 x 1.81) = .009). There were no significant correlations between PPT and accuracy in factive trials, τ = -.021, p = .445, or counterfactive trials, τ = .191, p = .111. WASI-II Matrices scores did not correlate significantly with accuracy in factive trials, τ = .164, p = .147. The correlation with accuracy in counterfactive trials was not significant below the adjusted threshold, τ = .284, p = .035. Brixton scores did not correlate with accuracy in factive trials, τ = .014, p = .464, or counterfactive trials, τ = .058, p = .356.

(*H7) Age in NBD controls*

We adjusted the significance threshold for the two outcome variables (p = .05 / 1.81 = .028). Age correlated significantly with accuracy in factive trials, τ = -.344, p = .009. It did not correlate significantly with accuracy in counterfactive trials, τ = -.152, p = .147.

(H8) Education in NBD controls

We adjusted the significance threshold for the two outcome variables (p = .05 / 1.81 = .028). Education did not correlate significantly with accuracy in factive trials, τ = .048, p = .4, or accuracy in counterfactive trials, τ = .077, p = .343.

In summary, older NBD controls were less accurate in factive interpretation trials. Participants with higher scores in the BNT were more accurate in counterfactive interpretation trials, with a trend in a similar direction for WASI-II Matrices.

2.3.3 Post-hoc analysis

Correlations between standardized tests and regression models

Appendix B presents all correlations, separately for each group. Significant correlates of SPM performance showed notable correlations with another. We list correlations with one-tailed p < .05, as one would expect positive correlations between test scores especially in the aphasic group, but advise general caution given the post-hoc nature of these tests. In the aphasic group, TROG-2 scores correlated with WASI-II Matrices scores, τ = .33, p = .028, the BNT, τ = .6, p < .001, and the PPT, τ =

.326, p = .034. WASI-II and PPT scores were correlated, τ = .37, p = .031. BNT and PPT scores were also correlated, τ = .3, p = .046.

On the basis of these correlations we explored the degree to which variables explained unique portions of the variance in factivity SPM accuracy, indicating their independence as explanatory variables. We conducted multiple linear regressions for this purpose. With our sample sizes, linear regressions do not have enough power to provide a generalizable model. However, R^2 values are reliable and do capture overlap within our sample. We computed two regressions, one for factive, one for counterfactive interpretation trials, and entered all variables which showed significant correlations with either trial type. We ranked variables to account for non-parametric distributions. We again report one-tailed p-values.

The model used a stepwise selection which starts with the strongest independent variable and only adds more if they significantly improve the model. In our case this means that if it selects only one independent variable, the portion explained by the others overlap. Within the aphasic group and factive trials, stepwise regressions produced a moderately strong model, F(1,18) = 11.811, p = .002, adjusted R^2 = .363. One variable was selected, TROG-2 scores, β = .629, p = .002. The model for accuracy in counterfactive trials also had moderate strength, F(1,18) = 11.382, p < .001, adjusted R^2 = .52. Two variables were selected, TROG-2 scores, β = .691, p = .003, and WASI-II Matrices scores, β = .335, p = .034.

Within the control group, significant correlates did not correlate with one another. We therefore assumed that explained variance would not overlap. We selected age, BNT scores and WASI-II Matrices scores. The model for accuracy in factive trials was weak, F(1,25) = 9.749, p = .002, adjusted R^2 = .252. Age was selected, β = -.53, p = .002. For accuracy in counterfactive trials the model was moderately strong, F(1,25) = 7.774, p = .001, adjusted R^2 = .343. The regression selected two variables, WASI-II Matrices scores, β = .467, p = .004, and BNT scores, β = .413, p = .008.

Errors in the SPM task

We analyzed what kind of errors participant groups made. For factive trials (e.g., *The man knows that it is warm outside*), we categorized foils as situation mismatch (e.g., selection of a picture in which it is cold outside) or experiencer mismatch (e.g., it is warm outside, but the man wears thick clothing). On average, controls selected .57 (SD = .97) situation mismatches and 1.07 (SD = 1.28) experiencer mismatches. A Wilcoxon test revealed a significant difference, Z = 3.217, p < .001, *r* = .41. Participants

with aphasia selected on average .9 (SD = 1.07) situation mismatches and 1.75 (SD = 2.1) experiencer mismatches. The difference was significant, Z = 2.379, p = .017, *r* = .37.

In counterfactive interpretation trials (e.g., *The man thinks that it is warm outside*), all foils were experiencer mismatches. We therefore categorized them as embedded clause match (e.g., it is warm outside and the man wears thick clothing) and embedded clause mismatch (e.g., it is cold outside and the man wears thick clothing). Controls on average selected 1 (SD = 1.43) embedded clause matches and .03 (SD = .18) embedded clause mismatches. The difference was significant, Z = 3.448, p = .001, *r* = .45. Participants with aphasia selected on average 4.05 (SD = 3.69) embedded clause matches and .55 (SD = .83) embedded clause mismatches. The difference was significant, Z = 3.536, p < .001, *r* = .56.

We performed chi-square analyses to find out if error distribution was different between groups. In factive trials, the control group selected 17 situation mismatches and 32 experiencer mismatches. For participants with aphasia, the numbers were 18 and 25, respectively (note that the aphasic group was smaller). The difference in error distribution was not significant, χ^2 (2, *N* = 102) = .006, p = .94. In counterfactive trials, the control group selected 30 foils with embedded clause matches and 1 foil with an embedded clause mismatch. In the aphasic group, the numbers were 81 and 11, respectively. The difference in error distribution was not significant, χ^2 (2, *N* = 123) = 2.007, p = .16.

3. Discussion

We designed a SPM test to investigate the ability of participants with aphasia and neurotypical controls to reach appropriate factive and counterfactive interpretations of clause embedding. We assumed that counterfactive interpretations pose higher demands on propositional and syntactic processing systems. Our first question concerned the overall performance of participants with aphasia across all sentence types, and as expected in a test using complex grammatical structures, aphasic participants performed worse than controls. Participants with aphasia did not perform worse in factive interpretation trials (which contained *know* and *it is clear* constructions), but showed a substantial and significant disadvantage in counterfactive interpretation trials (which used *think* and *it only seems* constructions). The data provide strong evidence for counterfactive interpretation being more demanding for people with aphasia. As for performance on factive interpretation trials, it is important to note that comparisons between aphasic and non-aphasic individuals depend much on aphasia severity, and that our group contained some individuals with very mild aphasia. We predict

that a more severely impaired group may show impairment in both trial types, albeit with greater difference for counterfactive trials.

Correlations with standardized test scores link errors in comprehension to syntactic, lexical, and non-verbal deficits. These results indicate the impact of severity of impairment at these levels and contribute to theories of factive and counterfactive processing (see introduction). With regards to syntax, aphasic performance in both factive and counterfactive trials was associated with sentence comprehension of a range of other structures, as measured by the TROG-2. Impaired sentence comprehension is one of the features that define aphasia, and our data support a syntactic hypothesis of factive and counterfactive interpretation in aphasia. Further research may aim to determine how these types of interpretations relate to comprehension of specific sentence types, for example other sentences with clause embedding.

Evidence for a lexical hypothesis was also present, albeit weaker. Aphasic participants with naming difficulties made significantly more mistakes in factive interpretation trials. In the control group, participants with lower naming scores made more mistakes in counterfactive trials. This relationship was found as a correlation between the BNT – a test of naming - , and SPM accuracy. However, word comprehension scores (CAT) did not correlate with SPM performance. This finding may be explained by test difficulty. The data show that aphasic participants performed closer to ceiling in the comprehension test. Also, variance in the BNT dataset is larger than in the CAT dataset, even relative to the test scale. We assume that the BNT is the more demanding probe of lexical capacity since naming involves word retrieval/recall, while comprehension tests measure recognition of picture-word matches. The actual lexical items may also be more difficult. For example, according to the British National Corpus (2007) the final item in the BNT (*abacus*) has a frequency of .05 per million words, while the least frequent word in CAT spoken word comprehension (*leek*) has a frequency of 1.4 per million. As a result, the CAT spoken word comprehension test is more likely to yield a ceiling effect (as evidenced in our data) which makes it less suited to pick up correlations. Note also that both BNT and the CAT subtest probe nouns/objects, while the critical word that drives factive and counterfactive interpretations in clausal embedding is the verb phrase. Verb processing is more likely to be affected in aphasia than noun processing (Druks, 2002), so future studies could include verb comprehension tests in participant profiling.

With regards to non-verbal behavior, the PPT (a non-linguistic test of picture semantics), was related to both factive and counterfactive interpretation in the aphasic group. Both tests require interpretation of picture material that goes beyond object recognition to the establishment of semantic relationships. Most striking however was the relationship between our comprehension task

and performance on WASI-II Matrices. In the aphasic group, performance on this non-verbal abstract reasoning test was the variable most strongly related to accuracy in counterfactive interpretation trials, but did not correlate significantly with accuracy in factive trials. We observed a similar pattern in the control group, although the correlation with counterfactive trials was above the adjusted significance threshold. The results may support the account that as clause embedding becomes non- or counterfactive, they become propositionally more complex and hence more demanding to a propositional system which can be affected by brain damage. Such a system may be relevant in matrix reasoning tests as these require, beyond visual perception, the ability to explore different solutions and integrate different variables or rules in order to develop explanations for matrix patterns. These may be in propositional form (e.g., the shape rotates; the box changes its color; this quantity increases) and may not be necessarily accessed using linguistic resources.

In factive, but not in counterfactive interpretation trials, older NBD participants were less accurate. While ageing has been associated with poorer auditory sentence comprehension, possibly as the result of changes to working memory capacity and hearing acuity (DeCaro, Peelle, Grossman, & Wingfield, 2016), this particular result is difficult to explain as we conceptualized counterfactive trials to be more difficult. We could not determine the exact nature of the ageing effect within our study.

Two individuals with aphasia were more accurate in counterfactive than factive interpretation trials. However, we advise caution since many SPM designs, including the design in this study, do not have enough trials for observations at case level to be reliable, especially if differences between performance across trial types are small. Longer SPM designs can start to explore the capacity to reach factive and counterfactive interpretations at an individual level, and may identify individuals with preserved counterfactive interpretation, but impaired factive interpretation capacities, similar to how SPM case studies identified individuals with impaired active, but intact passive voice comprehension (Druks & Marshall, 1995; Zimmerer, Dąbrowska, Romanowski, Blank, & Varley, 2014).

Duman et al. (2015), who tested aphasic comprehension of factuals and counterfactuals using SPM, suggest that impairment can come both from a disruption of networks for processing morphosyntactic information and networks responsible for “cognitive complexity”. Their evidence comes from what they describe as independent manipulations of morphosyntactic and cognitive complexity in their test stimuli. While our evidence comes from correlations between accuracy in our factivity and counterfactivity SPM task and other cognitive measures, it is compatible with this view. We found that scores for factivity comprehension, general syntactic comprehension and non-verbal reasoning were intercorrelated. However, regression models, while they should be taken with

caution within the framework of our study, suggest that the contributions of lexical, syntactic and non-verbal reasoning capacities to the model are mostly unique, resisting unification. Results are in line with models which see language and other cognitive networks as at least partially independent, such as strictly modular frameworks (Fodor, 1975, 2008; Pinker, 1994), or "supra-communicative" frameworks (Carruthers, 1998), which see networks as inherently independent, but regard internal verbalization as a way to coordinate subsystems and allocate additional memory resources to non-verbal thought. In both models, impairment of counterfactive interpretation in particular could be the result of disruption of language networks, networks for non-verbal reasoning, or both.

However, the relationship between factive and counterfactive interpretation of sentences and ToM needs to be explored more clearly, and while there is no evidence for ToM disruption in aphasia (see Introduction), not many individuals have been examined and it is possible that future studies find a stronger relationship between both levels, supporting theories that aim to unify language and cognition. One unified theory sees referential and propositional systems as essentially linguistic, and propositional thought to be grammatically structured (Hinzen & Sheehan, 2015; Sheehan & Hinzen, 2011).

Impairment to fluency has traditionally been associated with syntactic impairment. However, it has been challenged by accounts of heterogeneity within non-fluent groups (Berndt & Caramazza, 1999; Caramazza, Capitani, Rey, & Berndt, 2001; Tremblay & Dick, 2016), as well as reports of sentence comprehension deficits in people with Wernicke's type aphasia (Jefferies & Lambon Ralph, 2006; Ogar et al., 2011). While our study was not set up to examine this relationship, further investigation, either including fluency measures or systematically comparing fluent and non-fluent groups, can contribute to the debate. Given different possible sources of syntactic breakdown (e.g., lexical, syntactic, propositional), inclusion of additional cognitive probes might help discriminate between sub-types of sentence comprehension impairment, including the case where overall sentence comprehension scores are similar.

Further studies should explore the range of factivity constructions. We suggest testing processing of other classes of factives such as "true factives", which include *hate* and *regret*. Also, "non-factivity allows the embedded clause not only to be not true, as tested in our study, but to be true or unknown. This means that in a test of full comprehension of verbs like *think*, matching pictures should contain more than counterfactive contexts. Negation in factivity is another important phenomenon. In negated non-factives, there is no presupposition (e.g., *He doesn't think that it's warm outside*). This is not the case in negated factives (e.g., *He doesn't know that it's warm outside*).

We also need to be aware of general limitations of SPM tasks, such as their reliance on interpretation of static pictures and on multiple choice. SPM designs also do not control in any quantitative manner for visual complexity, which can occur at several levels (shapes, object depiction, object choice, constellation of scene). It is possible that unreported impairments of visual processing may impact test performance. We do not assume that this was the case in our study as aphasic performance was at ceiling on the visual PPT, but future study protocols could include additional assessments to control for such a factor. One important visual detail are facial expressions, which serve as cues to picture interpretation in the TROG-2 and in SPM components of the CAT. In our study, they were included to aid mental state interpretation in many, but not all, scenarios.

In summary, our data suggest that factive and counterfactive interpretations pose different cognitive demands, with the latter being more difficult for aphasic speakers. We suggest that counterfactives pose additional demands on propositional systems, and that these overlap with capacities involved in other verbal tasks, but also non-verbal reasoning tasks. We argue that factive, non-factive and counterfactive constructions pose a valuable opportunity to observe the relationship between lexical, syntactic and propositional cognition in aphasia. At the same time, research into brain injured populations can make its own contributions to understanding the interplay between language and propositional thought in all speakers. Our study also shows how these constructions can be tested using SPM, a paradigm commonly used not only for adult neurological populations, but also for children with developmental language disorder. Given their importance in communication and their relationship with ToM and propositional reasoning, our approach could result in a novel way to assess these aspects of cognition.

Authorship statement and acknowledgements

This study was conceived by WH, FD and VZ. VZ designed the experiment with input from WH, RV and FD. RV and VZ designed the protocol and recruited participants. VZ carried out the experiment and analyzed the data. VZ wrote the first draft of the manuscript. RV, WH and FD contributed to further drafts. This work was funded by the Arts and Humanities Research Council (AHRC), project AH/L004070/1. We thank our participants for their willingness to volunteer for this research. We also thank two anonymous reviewers for their useful comments and suggestions.

Our materials are available at https://www.cognitionandgrammar.net/s/Factivity-SPM.zip

References

Accepted for *Journal of Neurolinguistics* on 16/08/18

Apperly, I. A., Samson, D., Carroll, N., Hussain, S., & Humphreys, G. (2006). Intact first- and second-order false belief reasoning in a patient with severely impaired grammar. *Social Neuroscience*, *1*(3–4), 334–348. http://doi.org/https://doi.org/10.1080/17470910601038693

Baldo, J. V, Paulraj, S. R., Curran, B. C., & Dronkers, N. F. (2015). Impaired reasoning and problem-solving in individuals with language imapriment due to aphasia or language delay. *Frontiers in Psychology*, *6*, 1–14. http://doi.org/http://dx.doi.org/10.3389/fpsyg.2015.01523

Bánréti, Z., Hoffmann, I., & Vincze, V. (2016). Recursive subsystems in aphasia and Alzheimer's disease: Case studies in syntax and Theory of Mind. *Frontiers in Psychology*, *7*, 405. http://doi.org/10.3389/fpsyg.2016.00405

Berndt, R. S., & Caramazza, A. (1999). How "regular"□is sentence comprehension in Broca's aphasia? It depends on how you select the patients. *Brain and Language*, *67*(3), 242–247.

Bishop, D. (2003). *Test of Reception of Grammar (TROG-2) version 2*. London: Psychological Corporation.

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. http://doi.org/10.1016/j.neuroimage.2015.11.069

Burgess, P., & Shallice, T. (1997). *The Hayling and Brixton Tests*. Bury St. Edmunds, UK: Thames Valley Test Company.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal Theory of Mind. *Mind & Language*, *28*(5), 606–637. http://doi.org/10.1111/mila.12036

Caramazza, A., Capitani, E., Rey, A., & Berndt, R. S. (2001). Agrammatic Broca's aphasia is not associated with a single pattern of comprehension performance. *Brain and Language*, *76*(2), 158–184. http://doi.org/doi:10.1006/brln.1999.2275

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, *3*, 572–582. http://doi.org/https://doi.org/10.1016/0093-934X(76)90048-1

Carruthers, P. (1998). Thinking in language?: Evolution and a modularist possibility. In *Language and Thought* (pp. 94–119). Cambridge: Cambridge University Press.

DeCaro, R., Peelle, J. E., Grossman, M., & Wingfield, A. (2016). The two sides of sensory–cognitive interactions: Effects of age, hearing acuity, and working memory span on sentence comprehension. *Frontiers in Psychology*, *7*, 236. http://doi.org/10.3389/fpsyg.2016.00236

Druks, J. (2002). Verbs and nouns: a review of the literature. *Journal of Neurolinguistics*, *15*(3–5), 289–315. http://doi.org/https://doi.org/10.1016/S0911-6044(01)00029-X

Druks, J., & Marshall, J. C. (1995). When passives are easier than actives: two case studies of aphasic

comprehension. *Cognition*, *55*, 311–331. http://doi.org/https://doi.org/10.1016/0010-0277(94)00651-Z

Dudley, R., Orita, N., Hacquard, V., & Lidz, J. (2015). Three-year-olds' understanding of know and think (pp. 241–262). Springer International Publishing. http://doi.org/10.1007/978-3-319-07980-6_11

Duman, T. Y., Altınok, N., & Maviş, İ. (2016). Grammar and cognition: deficits comprehending counterfactuals in Turkish individuals with Broca's aphasia. *Aphasiology*, *30*(7), 841–861. http://doi.org/https://doi.org/10.1080/02687038.2015.1076926

Falmagne, R. J., Gonsalves, J., & Bennett-Lau, S. (1994). Children's linguistic intuitions about factive presuppositions. *Cognitive Development*, *9*(1), 1–22. http://doi.org/10.1016/0885-2014(94)90017-5

Fodor, J. A. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Friedmann, N., & Gvion, A. (2003). Sentence comprehension and working memory limitation in aphasia: A dissociation between semantic-syntactic and phonological reactivation. *Brain and Language*, *86*(1), 23–39. http://doi.org/10.1016/S0093-934X(02)00530-8

Gahl, S., & Menn, L. (2016). Usage-based approaches to aphasia. *Aphasiology*, 1–17. http://doi.org/10.1080/02687038.2016.1140120

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences*, *23*, 1–71. http://doi.org/10.1017/S0140525X00002399

Harris, R. J. (1975). Children's comprehension of complex sentences. *Journal of Experimental Child Psychology*, *19*(3), 420–433. http://doi.org/10.1016/0022-0965(75)90071-5

Hinzen, W., & Sheehan, M. (2015). *The Philosophy of Universal Grammar*. Oxford: Oxford University Press.

Howard, D., & Patterson, K. (1992). *Pyramids and Palm Trees: A test of semantic access from pictures and words.* Bury St. Edmunds, Suffolk: Thames Valley Test Company.

Howard, D., Swinburn, K., & Porter, G. (2004). *Comprehensive Aphasia Test*. Routledge: Psychology Press.

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. *Brain*, *129*(8), 2132–2147. http://doi.org/10.1093/brain/awl153

Johnson, C. N., & Maratsos, M. P. (1977). Early comprehension of mental verbs: Think and Know. *Child Development*, *48*(4), 1743. http://doi.org/10.2307/1128549

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test* (2nd ed.). Austin, TX: Pro-Ed.

Kay, J., Lesser, R., & Coltheart, M. (1997). *Psycholinguistic assessment of language processing in aphasia*. Hove, East Sussex, UK: Psychology Press.

Kiparsky, P., & Kiparsky, C. (1970). Fact. In M. Bierwisch & K. E. Heidolph (Eds.), *Progress in Linguistics: A Collection of Papers* (pp. 143–173). The Hague: Mouton.

Léger, C. (2007). The acquisition of two types of factive complements. In A. Gavarró & M. Freitas (Eds.), *Language acquisition and development: Proceedings of GALA* (pp. 337–347). Newcastle: Cambridge Scholars.

Lewis, S., Hacquard, V., & Lidz, J. (2012). The semantics and pragmatics of belief reports in preschoolers. In A. Chereches (Ed.), *Proceedings of SALT 22* (pp. 247–267).

McKenzie, D. (2012). Tools of the trade: A quick adjustment for multiple hypothesis testing. Retrieved February 8, 2017, from http://blogs.worldbank.org/impactevaluations/tools-of-the-trade-a-quick-adjustment-for-multiple-hypothesis-testing

Ogar, J. M., Baldo, J. V., Wilson, S. M., Brambati, S. M., Miller, B. L., Dronkers, N. F., & Gorno-Tempini, M. L. (2011). Semantic dementia and persisting Wernicke's aphasia: Linguistic and anatomical profiles. *Brain and Language*, *117*(1), 28–33. http://doi.org/10.1016/J.BANDL.2010.11.004

Peristeri, E., & Tsimpli, I.-M. (2013). Linguistic processing and executive control: Evidence for inhibition in Brica's aphasia. In N. Lavidas, T. Alexiou, & A.-M. Sougrari (Eds.), *Major Trends in Theoretical and Applied Linguistics* (Vol. 2, pp. 455–470). London: Versita.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, *316*(7139). http://doi.org/10.1136/bmj.316.7139.1236

Pinker, S. (1994). *The Language Instinct: The New Science of Language and Mind*. New York: Harper Collins.

Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association*, *62*(320), 1187–1200.

Raven, J., Raven, J., & Court, J. (2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Unionville, NY: Royal Fireworks Press.

Rubio-Fernandez, P., & Geurts, B. (2013). How to Pass the False-Belief Task Before Your Fourth Birthday. *Psychological Science*, *24*(1), 27–33. http://doi.org/10.1177/0956797612447819

Sankoh, A. J., Huque, M. F., & Dubey, S. D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, *16*(22), 2529–42. http://doi.org/10.1002/(SICI)1097-0258(19971130)16:223.0.CO;2-J

Sheehan, M., & Hinzen, W. (2011). Moving towards the edge. *Linguistic Analysis*, *3*(37), 405–458. http://doi.org/DOI:10.1023/A:1009780124314

Sloan Berndt, R., Mitchum, C. C., Haendiges, A. N., & Sandson, J. (1997). Verb Retrieval in Aphasia. 1.

Characterizing Single Word Impairments. *Brain and Language*, *56*(1), 68–106. http://doi.org/10.1006/brln.1997.1727

Swinney, D., & Zurif, E. B. (1995). Syntactic processing in aphasia. *Brain and Language*, *50*, 225–239.

The British National Corpus, version 2 (BNC XML Edition). (2007). Retrieved from http://www.natcorp.ox.ac.uk

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language*, *162*, 60–71. http://doi.org/10.1016/j.bandl.2016.08.004

Varley, R. (2014). Reason without much language. *Language Sciences*, *46*, 232–244. http://doi.org/10.1016/j.langsci.2014.06.012

Varley, R. A., Klessinger, N. J. C., Romanowski, C. A. J., & Siegal, M. (2005). Agrammatic but numerate. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(9), 3519–3524. http://doi.org/10.1073/pnas.0407470102

Varley, R. A., Siegal, M., & Want, S. C. (2001). Severe Impairment in Grammar Does Not Preclude Theory of Mind. *Neurocase*, *2001*(7), 489–493. http://doi.org/https://doi.org/10.1093/neucas/7.6.489

Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and "theory of mind" in an agrammatic aphasic patient. *Current Biology*, *10*(12), 723–726. http://doi.org/https://doi.org/10.1016/S0960-9822(00)00538-8

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence (WASI-II)*. San Antonio, TX: NCS Pearson.

Willems, R. M., Benn, Y., Hagoort, P., Toni, I., & Varley, R. (2011). Communicating without a functioning language system: Implications for the role of language in mentalizing. *Neuropsychologia*, *49*(11), 3130–3135. http://doi.org/10.1016/j.neuropsychologia.2011.07.023

Zimmerer, V. C., Cowell, P. E., & Varley, R. A. (2014). Artificial grammar learning in individuals with severe aphasia. *Neuropsychologia*, *53*, 25–38. http://doi.org/https://doi.org/10.1016/j.neuropsychologia.2013.10.014

Zimmerer, V. C., Dąbrowska, E., Romanowski, C. A. J., Blank, C., & Varley, R. A. (2014). Preservation of passive constructions in a patient with primary progressive aphasia. *Cortex*, *50*, 7–18. http://doi.org/https://doi.org/10.1016/j.cortex.2013.09.007

Zimmerer, V. C., & Varley, R. A. (2010). Recursion in severe agrammatism. In H. van der Hulst, J. Koster, & H. van Riemsdijk (Eds.), *Recursion and Human Language* (pp. 393–406). Berlin/New York: De Gruyter Mouton.

Accepted for *Journal of Neurolinguistics* on 16/08/18

Appendix A. Test protocol for all participants. Data were collected as part of the Language and Mental Health project which addresses a range of questions, many which go beyond the scope of this report. Therefore, not all data were analyzed for this report. Each session took approximately 90 minutes. Participants received 15 GBP for each session.

Test session 1	**Background interview; word-monitoring task; Boston Naming Test; Comprehensive Aphasia Test (sentence comprehension for both groups, word comprehension for participants with aphasia).**
Test session 2	Language elicitation; Pyramids and Palm Trees; PALPA 13 digit span; Raven's Colored Progressive Matrices; Factivity Sentence-Picture-Matching.
Test session 3	Brixton Spatial Anticipation Task; Grammaticality Judgment Task; Wechsler Abbreviated Spectrum of Intelligence; Test of Reception of Grammar II.

Appendix B. Relationships between performance measures, measured using one-tailed Kendall's tau correlations.

a) NBD controls

	Factives	**Non-factives**	**Age**	**Educ. (yrs)**	**TROG-2**	**PALPA 13**	**BNT**	**PPT**	**WASI-II**
Non-factives	τ = .12 p = .229								
Age	τ = -.34 p = .009	τ = -.152 p = .147							
Education (yrs)	τ = .08 p = .293	τ = .15 p = .165	τ = -.03 p = .42						
TROG-2	τ = .06 p = .364	τ = .13 p = .226	τ = -.31 p = .023	τ = .15 p = .17					
PALPA 13	τ = .19 p = .129	τ = .23 p = .082	τ = -.04 p = .404	τ = .01 p = .475	τ = .15 p = .191				
BNT	τ = .05 p = .381	τ = .33 p = .014	τ = -.128 p = .18	τ = .12 p = .197	τ = .05 p = .372	τ = -.08 p = .304			
PPT	τ = -.02 p = .445	τ = .19 p = .111	τ = .02 p = .441	τ = -.003 p = .492	τ = .13 p = .21	τ = -.26 p = .057	τ = .16 p = .139		
WASI-II Matrices	τ = .164 p = .147	τ = .28 p = .035	τ = -.07 p = .315	τ = .22 p = .072	τ = .28 p = .041	τ = .3 p = .019	τ = .01 p = .483	τ = .14 p = .175	
Brixton	τ = .01 p = .464	τ = -.06 p = .356	τ = -.1 p = .244	τ = .04 p = .406	τ = .3 p = .03	τ = .11 p = .246	τ = .04 p = .394	τ = -.11 p = .236	τ = .19 p = .104

b) Participants with aphasia

	Factives	**Non-factives**	**Age**	**Educ. (yrs)**	**TROG-2**	**PALPA 13**	**BNT**	**CAT**	**PPT**	**WASI-II**
Non-factives	τ = .5 p = .002									
Age	τ = -.04 p = .419	τ = -.12 p = .246								
Education (yrs)	τ = -.28 p = .06	τ = -25 p = .08	τ = .1 p = .285							
TROG-2	τ = .43 p = .008	τ = .429 p = .006	τ = -.07 p = .348	τ = -.14 p = .22						
PALPA 13	τ = .19 p = .155	τ = .14 p = .218	τ = .14 p = .219	τ = -.13 p = .245	τ = .34 p = .027					
BNT	τ = .43 p = .007	τ = .29 p = .041	τ = 0 p = .5	τ = -.21 p = .115	τ = .6 p < .001	τ = .34 p = .027				
CAT words	τ = .21 p = .127	τ = .164 p = .175	τ = .33 p = .027	τ = -.14 p = .22	τ = .18 p = .159	τ = -.03 p = .432	τ = .26 p = .063			
PPT	τ = .43 p = .01	τ = .4 p = .014	τ = -.07 p = .355	τ = -.46 p = .007	τ = .326 p = .034	τ = .05 p = .39	τ = .3 p = .046	τ = .11 p = .279		
WASI-II Matrices	τ = .23 p = .103	τ = .45 p = .004	τ = -.24 p = .079	τ = .11 p = .261	τ = .33 p = .028	τ = .06 p = .37	τ = .215 p = .1	τ = -.24 p = .085	τ = .37 p = .031	
Brixton	τ = .1 p = .294	τ = .1 p = .278	τ = -.17 p = .149	τ = .13 p = .221	τ = -.04 p = .397	τ = .04 p = .42	τ = .01 p = .487	τ = -.04 p = .408	τ = .4 p = .034	τ = .06 p = .371