Journal of Neurolinguistics 40 (2016) 98–111

Contents lists available at ScienceDirect

Journal of Neurolinguistics

journal homepage: www.elsevier.com/locate/jneuroling

Tracking sentence comprehension: Test-retest reliability in people with aphasia and unimpaired adults

Jennifer E. Mack [a, *], Andrew Zu-Sern Wei [a], Stephanie Gutierrez [a], Cynthia K. Thompson [a, b, c]

[a] Department of Communication Sciences and Disorders, Northwestern University, USA
[b] Cognitive Neurology and Alzheimer's Disease Center, Northwestern University, USA
[c] Department of Neurology, Northwestern University, USA

ARTICLE INFO

Article history:
Received 27 December 2015
Received in revised form 19 April 2016
Accepted 2 June 2016

Keywords:
Eyetracking
Aphasia
Sentence comprehension
Test-retest reliability

ABSTRACT

Purpose: Visual-world eyetracking is increasingly used to investigate online language processing in normal and language impaired listeners. Tracking changes in eye movements over time also may be useful for indexing language recovery in those with language impairments. Therefore, it is critical to determine the test-retest reliability of results obtained using this method.

Methods: Unimpaired young adults and people with aphasia took part in two eyetracking sessions spaced about one week apart. In each session, participants completed a sentence-picture matching task in which they listened to active and passive sentences (e.g., *The* [$_{N1+Aux}$ *woman was*] [$_{V}$ *visiting/visited*] [$_{NP/PP2}$ *(by) the man*]) and selected between two pictures with reversed thematic roles. We used intraclass correlations (ICCs) to examine the test-retest reliability of response measures (accuracy, reaction time (RT)) and online eye movements (i.e., the likelihood of fixating the target picture in each region of the sentence) in each participant group.

Results: In the unimpaired adults, accuracy was at ceiling (thus ICCs were not computed), with moderate ICCs for RT (i.e., 0.4–0.58) for passive sentences and low (<0.4) for actives. In individuals with aphasia, test-retest reliability was strong (0.59 < ICC < 0.75) for accuracy and excellent (>0.75) for RT for both sentence types. Similarly, for the unimpaired listeners, reliability of eye movements was moderate for passive sentences (NP/PP2 region) and low in all regions for active sentences. But, for the aphasic participant group, eye movement reliability was excellent for passive sentences (in the first second after sentence end) and strong for active sentences (V and NP/PP2 regions).

Conclusion: Results indicated moderate-to-low reliability for unimpaired listeners; however, reliable eye movement patterns were detected for processes specific to passive sentences (e.g., thematic reanalysis). In contrast, individuals with aphasia exhibited strong and stable performance across sentence types in response measures and online eye movements. These findings indicate that visual-world eyetracking provides a reliable measure of online sentence comprehension in aphasia, and thus may be useful for investigating sentence processing changes over time.

* Corresponding author. 2240 Campus Drive Evanston, IL 60208, USA.
E-mail address: Jennifer-mack-0@northwestern.edu (J.E. Mack).

http://dx.doi.org/10.1016/j.jneuroling.2016.06.001
0911-6044/

1. Introduction

In the last twenty years, visual-world eyetracking has become one of the most fruitful techniques for investigating language comprehension and production. Eyetracking has informed models of normal sentence processing at linguistic levels ranging from phonetic and phonological to lexical, sentence and discourse processing. In addition to the large eyetracking literature on language processing in monolingual adult speakers (see review in Huettig, Rommers, & Meyer, 2011), eyetracking has been used to examine the development of word and sentence comprehension processes in children (e.g. Borovsky, Elman, & Fernald, 2012; Borovsky, Sweeney, Elman, & Fernald, 2014; Fernald, Perfors, & Marchman, 2006; Mani & Huettig, 2012; Marchman & Fernald, 2008; Nation, Marshall, & Altmann, 2003) as well as the mechanisms of word comprehension in bilingual and second-language listeners (Blumenfeld & Marian, 2011; Chambers & Cooke, 2009).

In addition, eyetracking is increasingly used to characterize linguistic impairments in clinical populations. In the aphasia literature, eyetracking has played an important role in characterizing both lexical (Laurinavichyute, Ulicheva, Ivanova, Kuptsova, & Dragoy, 2014; Mirman & Graziano, 2012; Mirman, Yee, Blumstein, & Magnuson, 2011; Yee, Blumstein, & Sedivy, 2008) and sentence processing deficits (Bos, Hanne, Wartenburger, & Bastiaanse, 2014; Cho & Thompson, 2010; Choy & Thompson, 2010; Dickey & Thompson, 2009; Dickey, Choy, & Thompson, 2007; Hanne, Burchert, De Bleser, & Vasishth, 2015; Hanne, Sekerina, Vasishth, Burchert, & De Bleser, 2011; Lee & Thompson, 2011a, 2011b; Mack, Ji, & Thompson, 2013; Meyer, Mack, & Thompson, 2012; Patil, Hanne, Burchert, De Bleser, & Vasishth, 2015; Sheppard, Walenski, Love, & Shapiro, 2015; Thompson & Choy, 2009). For example, in one study (Meyer et al., 2012), we examined eye movements in unimpaired older adults and individuals with aphasia as they listened to active (e.g., *The man was visiting the woman*) and passive sentences (e.g., *The man was visited by the woman*) and selected between two pictures with reversed thematic roles. Unimpaired adults showed evidence of incremental agent-first (A1) processing, initially interpreting (non-case-marked, animate) subjects as agents (cf. Hanne et al., 2015). After presentation of the disambiguating verb morphology (*visit**ing** vs. visit**ed***), they rapidly fixated the target picture in both sentence types; in passive sentences, this likely reflects thematic reanalysis in which the initial A1 interpretation is revised (Hirotani, Makuuchi, Ruschemeyer, & Friederici, 2011; Mack, Meltzer-Asscher, Barbieri, & Thompson, 2013). The aphasic individuals, in contrast, did not show incremental A1 processing, fixating both pictures with equal frequency prior to verb offset. Following presentation of the disambiguating verb morphology, they exhibited delays in fixating the target picture in active sentences and never consistently did so in passive sentences (cf. Hanne et al., 2011). These findings highlight differences between unimpaired adults and aphasic listeners with respect to sentence processing latency and incremental interpretation.

Other populations in which eyetracking has been used to quantify performance patterns in language processing include individuals with apraxia of speech (Lee, Mirman, & Buxbaum, 2014), autism spectrum disorder (Brock, Norbury, Einav, & Nation, 2008; Mirman, Irwin, & Stephen, 2012; Venker, Eernisse, Saffran, & Ellis Weismer, 2013), and children and adolescents with language impairments (Borovsky, Burns, Elman, & Evans, 2013; McMurray, Munson, & Tomblin, 2014; McMurray, Samelson, Lee, & Tomblin, 2010; Nation et al., 2003). Eyetracking is also a potentially useful tool for measuring change over time, reflecting learning and/or language recovery in individuals with impaired language, though little research to date has used eyetracking for this purpose. In one study, Kim and Lemke (2016) used eyetracking-while-reading to test a participant with acquired alexia prior to administration of a text-based reading treatment program. After treatment, using the same task, the participant evinced more normal-like eye movements, reflecting facilitation of a lexical-semantic reading strategy.

Establishing the test-retest reliability of visual-world eyetracking is an important step in evaluating the usefulness of this method as a measure of performance patterns or language change over time. However, relatively little research has addressed this issue. One previous study (Farris-Trimble & McMurray, 2013) investigated test-retest reliability of measures of lexical access derived from visual-world eyetracking. In two test sessions, unimpaired young adult participants heard words (e.g., *horn*) while viewing arrays of four pictures including the target picture, a cohort competitor (e.g., *horse*), a rhyme competitor (e.g., *corn*), and an unrelated competitor (e.g., *box*). Multiple parameters of the time course of fixations to each picture type were modeled for each participant and test session. Correlation analyses comparing the time course parameters across test sessions indicated moderate-to-high reliability, especially for the timing of the rise in fixations to the target picture. In contrast with lexical access, the test-retest reliability of visual-world eyetracking in sentence comprehension tasks has not yet been investigated.

Only one previous study of which we are aware has examined test-retest reliability of sentence comprehension accuracy in unimpaired adults and individuals with aphasia, and no previous studies have examined RT or online eye movements. McNeil and colleagues tested reliability of performance on the Revised Token Test, which examines comprehension accuracy with sentences of varying length and complexity (McNeil et al., 2015). Their unimpaired adults showed modest reliability, likely due to near-ceiling performance on the task. However, the aphasic individuals showed excellent test-retest reliability, suggesting that sentence comprehension ability was reliable across test sessions. However, other studies have revealed a high degree of within-subject variability in performance in aphasia, e.g., across different sentence comprehension tasks (Caplan, Waters, Dede, Michaud, & Reddy, 2007) and – beyond the domain of sentence comprehension – across test sessions in narrative production (Boyle, 2014), confrontation naming (Freed, Marshall, & Chuhlantseff, 1996), and attentional tasks (Villard & Kiran, 2015).

In the present study, we examined the reliability of eye movements as participants performed a sentence-picture matching task (i.e., matching an auditorily-presented sentence to one of two pictures depicting possible interpretations of the sentence, as in Meyer et al., 2012). Measures of response accuracy and latency (RT) in this task have played an important

Table 1
Demographic and language testing measures for participants with aphasia.

Participant	Age	Gender	Education (years)	Months post-onset	WAB-R AQ	NNB N	NNB V	NAVS SPPT: C	NAVS SPPT: NC	NAVS SCT: C	NAVS SCT: NC	WPM	% GS
AP01	51	M	16	82	69.7	100.0%	100.0%	66.7%	0.0%	66.7%	66.7%	42.3	9.1%
AP02	35	F	19	58	83.7	100.0%	100.0%	80.0%	46.7%	66.7%	73.3%	54.4	93.3%
AP03	52	F	16	76	75.8	96.7%	100.0%	73.3%	46.7%	86.7%	53.3%	42.8	64.7%
AP04	53	F	13	103	53.5	96.7%	93.3%	80.0%	46.7%	93.3%	33.3%	36.4	0.0%
AP05	53	M	21	39	74.1	100.0%	100.0%	60.0%	0.0%	80.0%	26.7%	32.2	6.7%
AP06	41	M	16	16	89	100.0%	100.0%	80.0%	33.3%	86.7%	66.7%	120.0	78.1%
AP07	48	M	16	17	85	83.3%	86.7%	66.7%	13.3%	80.0%	40.0%	49.7	45.5%
AP08	29	F	19	28	84.1	100.0%	100.0%	93.3%	73.3%	93.3%	40.0%	n/a	n/a
AP09	41	M	16	85	76.2	100.0%	100.0%	80.0%	33.3%	86.7%	46.7%	26.0	50.0%
AP10	64	M	18	13	71.1	96.7%	100.0%	33.3%	0.0%	46.7%	46.7%	64.0	30.0%
AP11	46	M	18	34	53.5	80.0%	86.7%	33.3%	0.0%	80.0%	60.0%	81.4	58.3%
AP12	59	F	13	20	78.6	96.7%	93.3%	26.6%	20.0%	60.0%	73.3%	75.5	66.7%
Mean	**47.7**		**16.8**	**47.6**	**74.5**	**95.8%**	**96.7%**	**64.4%**	**26.1%**	**77.2%**	**52.2%**	**56.8**	**45.7%**
SD	**9.9**		**2.4**	**31.8**	**11.4**	**6.8%**	**5.3%**	**21.9%**	**24.4%**	**14.3%**	**15.8%**	**27.3**	**30.8%**

Notes: WAB-R AQ: Western Aphasia Battery-Revised, Aphasia Quotient; NNB: Northwestern Naming Battery, Auditory Comprehension subtest; N: Nouns; V: Verbs; NAVS: Northwestern Assessment of Verbs and Sentences; SPPT: Sentence Production Priming Test; SCT: Sentence Comprehension Test; C: Canonical Sentences; NC: Noncanonical sentences; WPM: Words Per Minute; % GS: % Grammatical sentences.

role in characterizing sentence comprehension deficits in aphasia (Berndt, Mitchum, & Haendiges, 1996; Caplan, Michaud, & Hufford, 2013; Caplan, Waters, & Hildebrandt, 1997; Caplan et al., 2007; Caramazza & Zurif, 1976; Grodzinsky, 1986; Grodzinsky, Piñango, Zurif, & Drai, 1999; Thompson et al., 2013). Sentence-picture matching has also been used, though less frequently, to investigate sentence comprehension processes in unimpaired adults (Malyutina & den Ouden, 2015; Pickering, McLean, & Branigan, 2013; Raffray & Pickering, 2010) and several recent studies have examined eye movements in both unimpaired and aphasic listeners using this task (Bos et al., 2014; Hanne et al., 2011, 2015; Meyer et al., 2012; Patil et al., 2015); see also Hochstadt (2009), who used this paradigm to investigate online sentence comprehension in individuals with Parkinson's Disease. This paradigm provides measurements of the likelihood of fixating the target picture at various time points as sentences unfold in real time. From this information, it is possible to investigate the magnitude and latency of participants' responses to linguistic information as it is presented, as well as incremental interpretations formed during sentence comprehension.

Based on these findings, we expected that RT and online eye movements might provide sensitive measures of sentence comprehension performance and thus reveal stable performance patterns. We anticipated high test-retest reliability in aphasic individuals for both offline and online measures, reflecting stable sentence comprehension ability and processing routines. However, we expected that impaired listeners might also show more within-subject (across session) variability as compared to unimpaired listeners, consistent with previous findings in the literature.

2. Methods

2.1. Participants

Twenty-one young adults, all native speakers of English, participated in the study (twelve women, nine men; mean age = 21). Two additional unimpaired individuals were tested but were excluded from the study due to poor eyetracker calibration. All were recruited from the undergraduate population at Northwestern University and reported no history of neurological, psychiatric, speech, language, or learning impairments. All participants passed a pure-tone audiometric screening prior to taking part in the experiment (threshold levels of 25 dB at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz), and had self-reported normal or corrected-to-normal vision. In addition, twelve people with aphasia participated in the study; one additional individual was tested but excluded due to poor eyetracker calibration. The individuals with aphasia were native speakers of English and passed a screening for visual acuity as well as a pure-tone audiometric screening (40 dB, 1000 Hz).[1] The study was approved by the Institutional Review Board at Northwestern University and all participants provided informed consent.

Table 1 summarizes the demographic and language testing information for the aphasic individuals included in the study. All were survivors of left-hemisphere strokes (single strokes in all cases except aphasic participant 07 (AP07), who had two strokes in rapid succession) and were in the chronic stage of recovery (at least one year post-stroke). They exhibited mild to moderately severe aphasia as measured by the Aphasia Quotient (AQ) from the Western Aphasia Battery-Revised (WAB-R) (Kertesz, 2006) and showed largely preserved word comprehension, as demonstrated by noun and verb comprehension scores on the Auditory Comprehension subtest of the Northwestern Naming Battery (Thompson & Weintraub, 2014). The

[1] The experimental stimuli were presented at a normal or slightly-louder-than-normal conversational volume (60–70 dB) and no participants reported difficulty in hearing the stimuli.

Sentence Production Priming Task (SPPT) and Sentence Comprehension Task (SCT) of the Northwestern Assessment of Verbs and Sentences (NAVS) (Thompson, 2011) were used to assess grammatical sentence production and comprehension ability; age-matched controls have previously been shown to perform at ceiling on these tasks (Cho-Reyes & Thompson, 2012). All aphasic participants showed impaired grammatical sentence production, with deficits ranging from relatively mild (e.g., AP08) to severe (e.g., AP10), and all produced canonical sentences more accurately than noncanonical sentences. Sentence comprehension was also impaired, and eight of 12 participants showed more accurate comprehension of canonical vs. noncanonical sentences. Further, linguistic analysis of a narrative language sample (Cinderella story; unavailable for AP08), showed that all participants except one (AP06) exhibited reduced fluency, defined as greater than two standard deviations below the mean words per minute (WPM) in a set of normative data from 13 age-matched controls (M = 132.2; SD = 18.8) (Thompson et al., 2012). All participants except one (AP02) also showed a reduced proportion of grammatical sentences, which was more than two standard deviations below the mean for the normative group of age-matched controls (M = 93.0%; SD = 4.4%) (Thompson et al., 2012). Thus, the aphasic participants exhibited language profiles generally consistent with mild-to-moderate agrammatism.

2.2. Materials

The experimental sentences consisted of active (e.g., *The man was visiting the woman*) and passive sentences (e.g., *The man was visited by the woman*). The sentences contained 24 semantically-reversible transitive verbs, all with regular past participles, each of which appeared once in the active and passive conditions, for a total of 48 critical trials. In order to minimize lexical processing demands, the sentences contained only four nouns referring to the agent and theme of the sentence: *man, woman, boy,* and *girl*. For half of the verbs, the identities of the agent and theme were reversed between active and passive sentences. Male and female participants were equally likely to be agents in both active and passive sentences. The sentences were recorded by a male native speaker of English at a normal speech rate (Mean = 4.0 syllables/second). Across conditions, the sentences were controlled for overall length in syllables and length of each noun phrase in ms (*p's* > 0.3).

For each verb, we developed a pair of action pictures (5 by 6 in.) with reversed semantic roles containing one male and one female participant (e.g., a woman visiting a man; a man visiting a woman), which were presented on a computer screen, placed four inches apart (see Fig. 1). In both active and passive conditions, the correct picture appeared equally often in the left and right boxes. Within pictures, the agent was also equally likely to appear on the left and the right sides of the scene.

There were also 24 filler trials, consisting of intransitive sentences (e.g., *The woman was dancing*) paired with pictures of a male and a female participant performing that action. The trials were presented in a pseudorandom order, with no more than three trials in a row from the same condition. Active and passive trials containing the same verb appeared at least 10 trials apart. In addition, the correct picture appeared in the same box for no more than five trials in a row.

2.3. Procedure

Participants' eye movements were recorded using an Applied Science Laboratories (ASL) 6000 remote eye tracker with a sampling rate of 60 Hz. In a dimly-lit room, participants were seated in front of a computer monitor and placed their chins on a chinrest, which was set at a height where the eyes would be level with the center of the monitor. Participants were instructed to listen to each sentence and click on the picture that matched the sentence. They were familiarized with the task through a five-item practice session. Each trial began with a fixation cross, which the participant clicked. Then, the picture was presented and remained on the screen as the sentence was presented auditorily, starting 500 ms later. When the participant clicked on one of the pictures, or after 10 s if no picture was selected, the pictured disappeared and the next trial began. The eye tracker was recalibrated every 10 trials. Each test session lasted approximately 20 min.

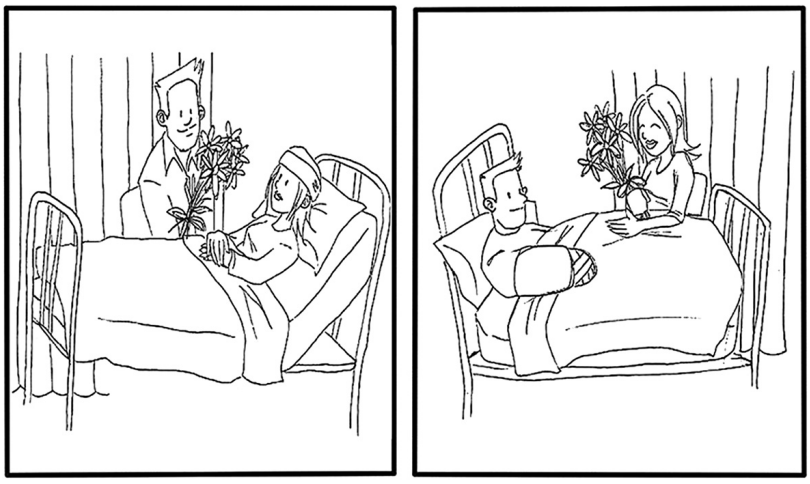

Fig. 1. Example visual stimulus.

The two test sessions were spaced approximately one week apart in the unimpaired participants (M (SD) = 7.3 (0.7) days). There was more variability in the testing schedule for aphasic participants (M (SD) = 4.9 (5.1) days), but the mean number of days between test sessions did not significantly differ between participant groups (two-sample *t*-test with unequal variances, $p > 0.1$).

2.4. Data analysis

2.4.1. Accuracy and reaction time data

To test for differences in performance across participant groups, sentence types, and test sessions, we used mixed-effects regression on the trial-level data (logistic regression for accuracy data, linear regression for RT data; *lme4* and *lmerTest* packages in R; Bates, Maechler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2015; R Core Team, 2015). Outlier RTs, which were defined as those greater than 3 standard deviations from the participant's conditional mean RT, were excluded (1.7% of the data in young adults and 1.1% in participants with aphasia). For both accuracy and RT, we first tested for group effects in the full data set (fixed effects: participant group, sentence type, test session, and their interactions), and then modeled the data for each group independently (fixed effects: sentence type, test session, and their interaction). By-participant intercepts and slopes (sentence type, test session, sentence type x test session) as well as by-item intercepts and slopes (sentence type) were included in all models.[2]

In order to examine the amount of variability between and within subjects (across test sessions) in each participant group, we computed coefficient of variation scores (COV; see e.g., Villard & Kiran, 2015). A COV is a ratio of the standard deviation of a data set to its mean. For each participant group and sentence type, we computed (1) between-subjects COV (BS_COV), by dividing the standard deviation of performance across participants (accuracy or RT, averaged across test sessions) by the mean performance across participants, and (2) within-subjects COV (WS_COV), by computing COVs for each individual (i.e., the standard deviation divided by the mean of their performance across the two test sessions), and averaging these values across the participant group. Because a single BS_COV value was generated for each group and sentence type, statistical analyses were precluded and the values were compared numerically across participant groups. WS_COVs were compared across participant groups and sentence types using mixed-effects ANOVAs (Type-III sum of squares with contrast-coded predictors, using the *ezANOVA* function of the *ez* package in R; Lawrence, 2013).

To quantify test-retest reliability, intraclass correlations (ICCs) were computed for each participant group and sentence type. Participants' mean accuracy and RT values from each session were entered into two-way random-effects consistency ICCs, using the *irr* package in R (Gamer, Lemon, Fellows, & Singh, 2012). ICCs describe the consistency between two or more raters or evaluation methods (in this case, test sessions) for the same set of participants, with an ICC of 1 reflecting perfect consistency (Shrout & Fleiss, 1979). Though there are no set criteria for the interpretation of ICC values, previous studies have interpreted values greater than approximately 0.75 as reflecting excellent reliability, between 0.59 and 0.75 as reflecting strong reliability, between 0.4 and 0.58 as reflecting moderate reliability, and values less than 0.4 as reflecting fair or low reliability (Bennett & Miller, 2010; Cicchetti & Sparrow, 1981; Farzin, Scaggs, Hervey, Berry-Kravis, & Hessl, 2011; Guo et al., 2012; Zanto, Pa, & Gazzaley, 2014). In keeping with this, we used these ICC values as a guide to interpretation of our data.

2.4.2. Eye movement data

The eye movement data were preprocessed using EYENAL (Applied Science Laboratories). Preprocessing consisted of assigning the eye movement data to fixations to areas of interest (i.e., pictures) in the visual array. A fixation was defined as a gaze to the same position on the screen within one degree of visual angle that lasted at least 100 ms. The quality of the eye movement data was assessed by computing, for each trial, the summed duration of fixations on the target and distractor pictures during the sentence. Trials in which the summed picture-fixation duration was greater than 50% of the sentence length were considered "high-quality" eye data.[3] The proportion of trials with high-quality data was compared across participant groups using a two-sample *t*-test.

Next, the time course of fixations on the target (correct) picture was computed for each participant, sentence type, and test session. Specifically, the data were aggregated into 50 ms windows time-locked to the offset of the verb, and for each window, we computed the proportion of trials in which the participant fixated the target picture, out of all trials in which the participant fixated one of the two pictures. Then, the eye movement curves were modeled using local polynomial regression (*loess* function in R), using quadratic polynomials with a span of 1 s. This was done in order to reduce bin-to-bin noise in the eye movement curves. See Fig. 2 for examples of raw eye movement data and modeled eye movement curves for one representative participant from each group. The modeled eye movement curves were averaged across individuals to produce group eye movement curves for each sentence type and test session.

The dependent variable for statistical analyses was the *target advantage* (i.e., the strength of the tendency to fixate on the target vs. the distractor picture) in each region of the sentence. The sentence regions were as follows: the subject noun and subsequent auxiliary (N1 + Aux; e.g., *man was*; mean length = 646 ms), the verb (V; e.g., *visiting/visited*; mean

[2] By-item random slopes were removed in the model of the control RT data to enable model convergence.

[3] The time spent *not* fixating either picture includes fixations outside the picture boundaries, saccades, blinks, and transient data loss, all of which are common even in high-quality data.

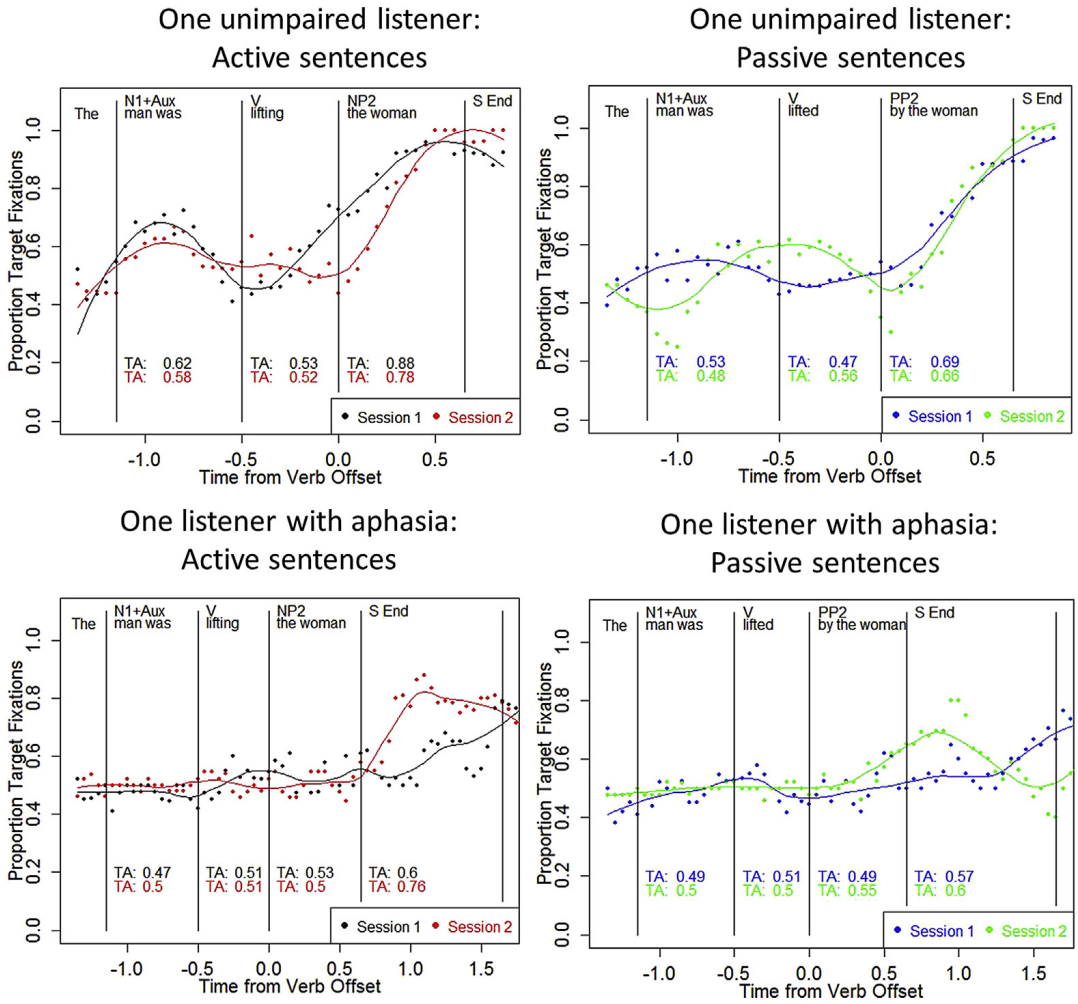

Fig. 2. Eye movement data from one representative unimpaired participant and one representative person with aphasia. The dots indicate raw eye movement data points, i.e., the proportion of trials in which the participant was fixating the target picture, aggregated into 50 ms bins. The lines indicate the eye movement curves as modeled through local polynomial regression. TA = target advantage (i.e., the area under the eye movement curve in each region, divided by the length of the region).

length = 513 ms), and the postverbal noun phrase/prepositional phrase (NP/PP2; e.g., *(by) the woman*; mean length = 683 ms). YA participants tended to respond quite quickly after sentence end (group mean RT = 388 ms; mean RT of the fastest individual = 209 ms), and thus we did not statistically examine eye movements after sentence offset. For participants with aphasia, who showed longer RTs (group mean RT = 2765 ms; mean RT of the fastest individual = 1009 ms), we also computed the target advantage in the first second after the end of the sentence (S End region).

Target advantage scores were computed by calculating the area under the modeled eye movement curve in each sentence region, divided by the mean length in seconds of that region. A score greater than 0.5 indicates a tendency to fixate the target picture (max score = 1, a 100% tendency to fixate the target picture) and a score less than 0.5 a tendency to fixate the distractor picture (min score = 0, a 0% tendency to fixate the target picture). See Fig. 2 for an illustration of target advantage scores for one unimpaired and one aphasic individual. This measure reflects both the latency and magnitude of eye movement effects. All else being equal, effects that emerge early in a window and are sustained throughout that window have a larger target advantage score than late-emerging effects. Similarly, effects of larger magnitude have larger target advantage scores than effects of smaller magnitude. Though area-under-curve computations are not commonly used in visual-world eye-tracking, they are often used in analyses of other time course data, such as event-related potentials (ERPs) (Luck, 2014).

To test for effects of participant group, sentence type, and test session on eye movement patterns (i.e., target advantage scores), we used repeated-measures ANOVAs (Type-III sum of squares with contrast-coded predictors, using the *ezANOVA* function of the *ez* package in R; Lawrence, 2013).[4] Separate analyses were performed for each sentence region. We first tested for effects of group by analyzing both groups' data together (independent variables: group, sentence type, test session, and

[4] ANOVA was used instead of mixed-effects regression because the eye movement data were aggregated at the subject level. In these analyses, 0.5 was subtracted from all target advantage scores (Intercept = 0, an equal tendency to fixate the target and distractor pictures).

Table 2
Accuracy and RT data: Performance by session in unimpaired and aphasic listeners.

	Accuracy		RT (seconds)	
	Unimpaired	Aphasia	Unimpaired	Aphasia
Active sentences				
Session 1				
Mean	99.8%	74.0%	0.434	2.780
SD	0.9%	18.9%	0.163	1.340
Session 2				
Mean	99.8%	75.7%	0.299	2.645
SD	0.9%	16.7%	0.096	0.989
Passive sentences				
Session 1				
Mean	99.6%	55.2%	0.475	2.876
SD	1.3%	18.0%	0.166	1.308
Session 2				
Mean	99.0%	61.1%	0.342	2.758
SD	1.9%	19.2%	0.143	1.112

their interactions), and then performed separate analyses by group (independent variables: sentence type, test session, and their interaction).

To examine between- and within-subjects variability, we computed BS_COV and WS_COV values for each participant group, sentence type, and sentence region, using the same methods as described above. In order to examine test-retest reliability of eye movement patterns, the target advantage scores from each test session were entered into two-way random-effects consistency ICCs. ICCs were computed for each participant group, sentence type, and sentence region.

3. Results

3.1. Accuracy and RT

Table 2 presents each participant group's accuracy and RT for each sentence type and test session. The results of mixed-effects models of the accuracy and RT data appear in Table 3, and measures of between- and within-subject variability and test-retest reliability appear in Table 4. Due to a data collection error, accuracy and RT data were unavailable for one unimpaired participant. Starting with the accuracy data, mixed-effects regression models of the full data set showed that the young unimpaired group performed more accurately than the aphasic participant group ($z = -10.504$, $p < 0.001$) and overall, participants responded more accurately to active than passive sentences ($z = 2.383$, $p < 0.05$), with no other significant effects. Due to at-ceiling performance (M > 99%), the unimpaired participants' accuracy data were not modeled separately and no test-retest reliability measures were calculated. Between- and within-subjects variability measures for accuracy in this group were extremely low across sentence types (BS_COVs < 0.02; WS_COVs < 0.01).

Although the aphasic group showed poorer accuracy than the unimpaired participant group, they evinced significantly better accuracy for active than passive sentences ($z = 2.865$, $p < 0.01$), with no significant effects of test session or interaction between sentence type and test session, and test-retest reliability for this group was in the "strong" range for both sentence types (active: ICC = 0.641, $p < 0.01$; passive: ICC = 0.712, $p < 0.01$). The aphasic listeners also showed substantially greater between- and within-subject variability than unimpaired listeners (BS_COVs > 0.2, WS_COVs > 0.1, respectively). Comparing the within-subject variability in the aphasic and unimpaired groups using ANOVA showed a main effect of group ($F = 46.5$, $p < 0.001$), reflecting greater variability for aphasic than unimpaired listeners, as well as a main effect of sentence type ($F = 4.5$, $p < 0.05$), indicating greater variability for passive than active sentences.

Moving to the RT results, the unimpaired participants responded more rapidly than the aphasic participants ($t = 9.456$, $p < 0.001$) and there was an overall trend towards faster RTs in the second test session ($t = 1.988$, $p = 0.056$), with no other significant effects. The unimpaired group responded significantly more quickly in session 2 than in session 1 ($z = 4.655$, $p < 0.001$), but showed no effect of sentence type or interaction between test session and sentence type. The aphasic group showed no significant effects of test session, sentence type, or interaction between test session and sentence type. Between-subject variability was numerically higher in the aphasic group (BS_COV active = 0.409; passive = 0.422) than in the unimpaired group (BS_COV active = 0.271; passive = 0.325), but within-subject variability was higher in the unimpaired group (WS_COV active = 0.322; passive = 0.295) than the aphasic group (WS_COV active = 0.167; passive = 0.116). ANOVA of within-subject variability revealed a main effect of group ($F = 20.7$, p < 0.001), i.e., greater within-subject variability for unimpaired than aphasic listeners, but no effects of sentence type. For unimpaired listeners, RTs exhibited low test-retest reliability for active sentences (ICC = 0.101, $p > 0.1$) and moderate reliability for passive sentences (ICC = 0.475, $p < 0.05$). In contrast, the RTs of the aphasic participants showed excellent reliability for both active (ICC = 0.773, $p < 0.001$) and passive sentences (ICC = 0.915, $p < 0.001$).

Table 3
Mixed-effects regression models of the accuracy and RT data.

	Accuracy		RT	
	z	*p*	*t*	*p*
Unimpaired adults vs. Participants with aphasia				
Intercept	13.632	<0.001	12.491	<0.001
Group	−10.504	<0.001	9.456	<0.001
Sentence type	2.383	0.017	−1.247	0.2228
Session	0.300	0.764	1.988	0.0559
Group * sentence type	−0.472	0.637	−0.609	0.5466
Group * session	−0.766	0.444	−0.065	0.9488
Sentence type * session	−0.370	0.711	0.09	0.9284
Group * sentence type * session	0.595	0.552	0.05	0.9604
Unimpaired adults				
Intercept	n/a; at-ceiling performance		16.337	<0.001
Sentence type			−1.589	0.128
Session			4.655	<0.001
Sentence type * session			0.052	0.959
Participants with aphasia				
Intercept	4.295	<0.001	8.468	<0.001
Sentence type	2.865	0.004	−0.739	0.472
Session	−1.252	0.211	0.763	0.461
Sentence type * session	0.501	0.617	0.059	0.954

Notes: Reference levels are as follows: Group = unimpaired adults; Session = Session 1; Sentence Type = Passive.

Table 4
Accuracy and RT data: Between- and within-subject variability and test-retest reliability.

	Accuracy		RT (seconds)	
	Unimpaired	Aphasia	Unimpaired	Aphasia
Active sentences				
BS_COV	0.006	0.216	0.271	0.409
WS_COV	0.003	0.110	0.322	0.167
ICC	N/A	0.641	0.101	0.773
		p = 0.009	p = 0.320	p < 0.001
Passive sentences				
BS_COV	0.012	0.297	0.325	0.422
WS_COV	0.008	0.167	0.295	0.116
ICC	N/A	0.712	0.475	0.915
		p = 0.003	p = 0.015	p < 0.001

Note: BS_COV = between-subjects coefficient of variance; WS_COV = within-subjects coefficient of variance; ICC = intraclass correlation coefficient.

3.2. Eye movements

Eye data quality was strong in both participant groups: the unimpaired listeners had high-quality eye data in an average of 88.3% of trials (SD = 10.4%) and the individuals with aphasia in 79.4% of trials (SD = 14.6%). There was no significant difference between groups with respect to eye data quality (*t*-test, $p > 0.1$).

Fig. 3 illustrates the mean proportion of target fixations over the course of the sentence in each participant group. Table 5 summarizes the target advantage scores in each sentence region for each participant group, sentence type, and test session. The ANOVA results, which examined eye movements across sentence types and test sessions in each participant group, appear in Table 6. In the N1 + Aux and Verb regions, unimpaired participants exhibited agent-first processing, i.e., a tendency to fixate the picture in which the subject noun is the agent (the target picture in active sentences and the distractor picture in passive sentences), but the aphasic listeners did not. In these regions, there was an overall effect of sentence type (N1 + Aux region: $F = 12.398$, $p = 0.001$; V region: $F = 27.009$; $p < 0.001$) which was driven by the unimpaired group, as indicated by (1) interactions between group and sentence type in the full models of the data (N1 + Aux region: $F = 7.978$, $p < 0.01$; V region: $F = 14.819$; $p = 0.001$); (2) significant main effects of sentence type in the unimpaired group, indicating larger target advantages in active than passive sentences (N1 + Aux region: $F = 24.096$, $p < 0.001$; V region: $F = 61.681$; $p < 0.001$), and (3) no effects of sentence type in the aphasic group (p's > 0.1). There were no other significant effects in the N1 + Aux or V regions.

In the NP/PP2 region, the unimpaired group made more target fixations overall than the aphasic group ($F = 91.170$, $p < 0.001$), and participants overall made more target fixations for active than passive sentences ($F = 17.382$, $p < 0.001$), an effect that was driven primarily by the unimpaired group ($F = 34.602$, $p < 0.001$) as the aphasic listeners did not show a significant effect of sentence type ($p > 0.1$). In addition, there was a significant effect of test session ($F = 5.559$, $p < 0.05$) which was driven by the unimpaired listeners who made more target fixations in the second compared to the first test session ($F = 7.600$, $p < 0.05$); the aphasic group did not show an effect of test session in this region ($p > 0.1$). In the S End region (the

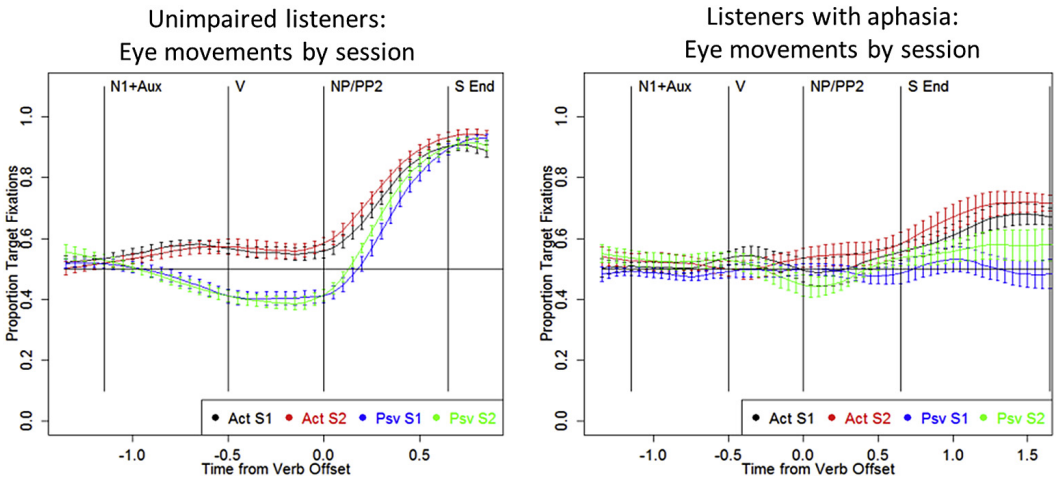

Fig. 3. Eye movements from each participant group, by test session and sentence type. Act = Active; Psv = Passive; S1 = Test Session 1; S2 = Test Session 2.

first second after sentence end), which was only analyzed for aphasic listeners due to rapid RTs in the unimpaired group, the aphasic group made more target fixations for active than passive sentences ($F = 12.192$, $p < 0.01$), and in the second compared to the first test session ($F = 10.920$, $p < 0.01$), but showed no interaction between sentence type and test session ($p > 0.1$).

Table 7 provides measures of between- and within-subject variability and test-retest reliability. The unimpaired and aphasic listeners exhibited numerically similar between-subject and within-subject variability in eye movements (range of BS_COV for young adults = 0.079–0.175; participants with aphasia = 0.087–0.250; range of WS_COV for young adults: 0.073–0.157; participants with aphasia = 0.073–0.119), and the ANOVA analyses of within-subject variability did not reveal any significant main effects of group or sentence type (p's > 0.1). For unimpaired listeners, test-retest reliability for active sentences was low in all sentence regions (ICC's < 0.4, p's > 0.1). However, for passive sentences, test-retest reliability was low only prior to verb offset (ICC's < 0.4, p's > 0.1) but in the "moderate" range during the NP/PP2 region (ICC = 0.483, $p < 0.05$). For aphasic listeners, in active sentences, test-retest reliability was low during N1 + Aux (ICC <. 4, $p > 0.1$), strong during the verb (ICC = 0.601, $p < 0.05$) and NP/PP2 (ICC = 0.646, $p < 0.01$), and moderate in the S End region (ICC = 0.462, $p = 0.056$). In passive sentences, test-retest reliability was poor during N1 + Aux (ICC = 0.01, $p > 0.1$), moderate during the verb (ICC = 0.464, $p = 0.055$) and NP/PP2 (ICC = 0.52, $p < 0.05$), and excellent in the S End region (ICC = 0.937, $p < 0.001$).

4. Discussion

The present study evaluated the test-retest reliability of visual-world eyetracking patterns during online sentence comprehension. Although one previous study investigated the reliability of the time course of lexical access as measured with eyetracking in healthy young adults (Farris-Trimble & McMurray, 2013), no previous studies have done so in the domain of sentence comprehension. Further, to our knowledge, no previous studies have addressed the test-retest reliability of eye-tracking in individuals with aphasia, despite the substantial contributions that this method has made to the understanding of language processing impairments in aphasia. In the present study, healthy young adult participants and people with aphasia performed a sentence-picture matching task in two sessions spaced approximately one week apart (paradigm adapted from Meyer et al., 2012). Within each participant group, we tested response stability, as measured by accuracy and reaction times (RTs), as well as online eye movements.

The young unimpaired participants performed with very high accuracy (M > 99%) and rapid RTs (M = 0.388 s), with no overall differences in accuracy or RT between active and passive sentences. Participants showed similar eye movement patterns to those reported by Meyer et al. (2012) for healthy older adults, and these patterns were qualitatively very similar across test sessions at the group level. The unimpaired group exhibited agent-first processing, i.e., a tendency to fixate the picture in which the subject noun is the agent, starting soon after the onset of the subject noun and persisting through presentation of the verb. This eye movement pattern is consistent with several previous studies which have reported agent-first processing in unimpaired adults (Hanne et al., 2015; Kamide, Scheepers, & Altmann, 2003; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Meyer et al., 2012). After presentation of the disambiguating verbal morphology (*visit**ing*** vs. *visit**ed***), participants fixated the target picture, reflecting rapid incremental integration of morphosyntactic information and, in the case of passive sentences, thematic reanalysis of the initial agent-first interpretation.

In the young unimpaired group, RT and eye movement measures exhibited similar test-retest reliability. The reliability of RTs was low for active sentences, but higher, in the moderate range, for passive sentences. For eye movements, the reliability of the target advantage was low for both sentence types during processing of the subject and verb, and remained low during the post-verbal NP/PP for active sentences, but was in the moderate range for passive sentences. Thus, the results provided useful information about the relative level of reliability of particular sentence comprehension processes. The low ICC values

Table 5
Eye movements (target advantage scores): Performance by session in unimpaired and aphasic participants.

	Unimpaired			Aphasia			
	N1 + Aux	V	NP/PP2	N1 + Aux	V	NP/PP2	S end
Active							
Session 1							
Mean	0.564	0.555	0.742	0.509	0.532	0.514	0.637
SD	0.063	0.081	0.079	0.049	0.095	0.122	0.108
Session 2							
Mean	0.550	0.565	0.777	0.519	0.509	0.552	0.685
SD	0.070	0.114	0.073	0.080	0.100	0.119	0.121
Passive							
Session 1							
Mean	0.480	0.405	0.648	0.483	0.496	0.488	0.507
SD	0.078	0.097	0.093	0.036	0.069	0.071	0.137
Session 2							
Mean	0.475	0.395	0.684	0.527	0.497	0.481	0.566
SD	0.059	0.084	0.071	0.081	0.108	0.082	0.136

Table 6
ANOVA models of the eye movement data.

Sentence region	N1 + Aux		V		NP/PP2		S end	
	F	*p*	*F*	*p*	*F*	*p*	*F*	*p*
Unimpaired adults vs. Participants with aphasia								
Intercept	5.665	0.024	0.309	0.582	107.266	<0.001	n/a: not analyzed	
Group	0.451	0.507	1.911	0.177	91.170	<0.001	for young adult	
Sentence type	12.398	0.001	27.009	<0.001	17.382	<0.001	group	
Session	0.518	0.477	0.143	0.708	5.559	0.025		
Group * sentence type	7.978	0.008	14.819	0.001	1.786	0.191		
Group * session	2.145	0.153	0.153	0.698	0.865	0.360		
Sentence type * session	0.762	0.389	0.007	0.934	0.875	0.357		
Group * sentence type * session	0.275	0.604	0.490	0.489	0.931	0.342		
Unimpaired adults								
Intercept	6.965	0.016	2.933	0.102	314.988	<0.001	n/a: not analyzed	
Sentence type	24.096	<0.001	61.681	<0.001	34.602	<0.001	for young adult	
Session	0.464	0.503	<0.001	0.992	7.600	0.012	group	
Sentence type * session	0.068	0.797	0.213	0.649	0.001	0.981		
Participants with aphasia								
Intercept	1.000	0.339	0.221	0.648	0.206	0.659	11.671	0.006
Sentence type	0.261	0.619	0.615	0.449	1.829	0.203	12.192	0.005
Session	1.417	0.259	0.352	0.565	0.770	0.399	10.920	0.007
Sentence type *session	1.327	0.274	0.409	0.536	1.488	0.248	0.086	0.775

obtained during the N1 and V regions of both sentence types suggest considerable within-subject variability in the magnitude and/or timing of agent-first processing effects (see e.g., the unimpaired participant in Fig. 2, who shows distinct eye movement patterns between test sessions in the N1 and V regions, especially for passive sentences). In contrast, relatively stable individual performance patterns were found during the post-verbal (PP2) region of passive sentences, during which thematic reanalysis takes place upon processing of the verb morphology. These patterns suggest that the sentence-picture matching paradigm may be used to capture important processes associated with sentence comprehension in unimpaired listeners. In future research, the paradigm could be used with more complex stimuli (e.g., requiring reanalysis) in order to maximize the chances of capturing stable individual performance patterns in unimpaired listeners.

However, the overall moderate-to-low reliability found for the unimpaired group contrasts with the results of Farris-Trimble and McMurray (2013), who noted moderate-to-strong reliability for eye movement measures during a lexical access task. There are several factors that may have contributed to this. First, the relatively modest number of experimental trials in the present study (n = 24 per participant, condition and test session), which was selected due to practical constraints on testing for aphasic participants, may have reduced the ability to detect subtle individual differences. Notably, the eye movement parameters from Farris-Trimble and McMurray (2013) were derived from as many as 640 trials per participant and test session. Second, and relatedly, due to the relatively modest number of trials, we chose to compute a single eye movement parameter for each sentence region (target advantage, i.e., the mean area under the eye movement curve), whereas Farris-Trimble and McMurray (2013) computed several parameters, which showed varying degrees of reliability. Thus, in order to detect stable eye movement patterns in unimpaired adults, it may be advisable to include a large number of trials to enable modeling of different parameters of eye movement curves. Third, the sentence comprehension task that we employed may have encouraged participants to use online processing strategies (e.g., agent-first processing) that do not take place in a lexical

Table 7
Eye movements (target advantage scores): Between- and within-subject variability, and test-retest reliability.

	Unimpaired			Aphasia			
	N1 + Aux	V	NP/PP2	N1 + Aux	V	NP/PP2	S end
Active							
BS_COV	0.094	0.130	0.079	0.087	0.167	0.205	0.149
WS_COV	0.090	0.127	0.073	0.105	0.094	0.119	0.109
ICC	0.228	0.071	0.235	−0.093	0.601	0.646	0.462
P	0.153	0.376	0.146	0.619	0.015	0.008	0.056
Passive							
BS_COV	0.089	0.175	0.107	0.088	0.156	0.138	0.250
WS_COV	0.130	0.157	0.083	0.073	0.092	0.088	0.092
ICC	−0.252	0.185	0.483	0.01	0.464	0.52	0.937
P	0.871	0.205	0.011	0.488	0.055	0.034	<0.001

Notes: BS_COV = between-subjects coefficient of variation; WS_COV = within-subjects coefficient of variation; ICC = intraclass correlation coefficient.

access task. Participants may have varied in their use of these strategies across test sessions, thereby reducing measures of test-retest reliability. Notably, the aphasic participants, who did not employ an online agent-first processing strategy, showed relatively strong eye movement reliability. Further research is needed to determine how language processing strategies, and change in these strategies over time, might affect the reliability of measures of normal sentence processing.

As compared to the young unimpaired group, the aphasic group performed less accurately (Overall M = 66.5%) and more slowly (Overall M RT = 2.765 s), with substantial individual variability in performance. Passive sentences were comprehended less accurately than active sentences, consistent with a large literature on comprehension of these sentence types in participants with aphasia, particularly those with agrammatism (Bastiaanse & Edwards, 2004; Burchert & De Bleser, 2004; Caplan et al., 1997; Grodzinsky et al., 1999; Meyer et al., 2012; Thompson et al., 2013). Consistent with the findings of Meyer et al. (2012) as well as Hanne et al. (2015), the aphasic individuals did not show online agent-first processing. Thus, these findings provide further support for the idea that aphasic listeners do not use agent-first strategies for sentence interpretation, at least not online (Meyer et al., 2012; in contrast with accounts such as the Trace Deletion Hypothesis (Grodzinsky, 1986) which assume agent-first interpretive strategies). There was no difference in target advantage scores between passives and actives during sentence presentation. However, after sentence offset (within 1000 ms in the S End region), participants exhibited a tendency to fixate the target picture, which was stronger in active than passive sentences. These findings are consistent with previous studies in which aphasic listeners show delayed fixations to the target picture in canonical structures and inconsistent (or at-chance) fixations of the target picture in noncanonical structures (Hanne et al., 2011; Meyer et al., 2012).

In individuals with aphasia, test-retest reliability was generally stronger than in unimpaired listeners, and was similar for response measures (accuracy, RT) and online eye movements. For both active and passive sentences, ICC values were strong for accuracy and excellent for RTs, indicating that the task captured stable response patterns. However, the reliability of eye movements showed differing patterns by sentence type and region. For both sentence types reliability was low during the N1 + Aux region; strong during the verb and NP/PP2 regions for actives, but moderate for passives; and moderate in the S End region for actives, yet excellent for passive structures. In other words, eye movements in active sentences were most reliable during the sentence, whereas eye movements for passive sentences were most reliable after sentence end. These generally strong test-retest reliability effects observed for the aphasic participants likely relate to the properties of the experimental paradigm as well as the participant group. The experimental paradigm was designed to minimize lexical processing demands and thus isolate grammatical sentence comprehension processes to the extent possible. The aphasic group, in general, showed relatively intact word comprehension ability but grammatical deficits of varying levels of severity that were evident in both sentence production and comprehension tasks. Thus, the experimental paradigm may have been especially well-suited for detecting stable patterns in this group.

Notably, the unimpaired adults in the present study were younger than the participants with aphasia, raising the possibility that age differences may have contributed to group differences in reliability. However, as mentioned above, the unimpaired young adult participants in the present study showed accuracy and eye movement patterns that were quite similar to those of unimpaired older adults in our previous study (Meyer et al., 2012). Therefore, we attribute between-group differences found in the present study to the presence/absence of aphasia, rather than age differences. However, little is known about how (or if) the reliability of sentence processes changes with age. Further research is needed to clarify this issue.

In addition to examining test-retest reliability, the present study computed measures of between- and within-subject (across session) variability in young unimpaired adults and those with aphasia. It is often noted anecdotally, by individuals with aphasia as well as clinicians and researchers, that considerable variability in language performance across days is common, and several studies have lent quantitative support to this idea (Boyle, 2014; Caplan et al., 2007; Freed et al., 1996; Villard & Kiran, 2015). Accordingly, we were particularly interested in the relative levels of within-subject variability in aphasic individuals as compared to unimpaired controls. For accuracy, the aphasic individuals showed more within-subject variability than unimpaired listeners, who performed nearly perfectly across sessions. However, for RTs, within-subject variability was actually less in the aphasic compared to the unimpaired group, likely due to the existence of significant

practice effects in the latter group (see following paragraph). Further, with respect to eye movements, no significant differences in within-subject variability were observed between groups. Thus, the present results suggest that high levels of day-to-day variability are not apparent for all people with aphasia and for all tasks. Specifically, for measures of the latency of sentence comprehension and online sentence comprehension processes, aphasic individuals may not show greater variability than unimpaired adults. However, further research is needed to completely understand variability in language processing in aphasia and the variables that affect it.

Notably, practice effects were seen in both participant groups. The unimpaired listeners responded more rapidly in the second test session, and also made more target fixations during NP/PP2. Although the aphasic individuals exhibited no significant changes across sessions in accuracy or RT, they made more target fixations during the S End region in the second test session. These effects are likely due to familiarity with the experimental task and stimuli. These findings are both instructive and encouraging for researchers with interest in using visual-world eyetracking to measure changes in individual eye movement patterns over time, e.g., in response to treatment for language impairments. They demonstrate that practice effects can be seen in eye movement data, and thus motivate the inclusion of multiple eyetracking baseline sessions in studies that will use this technique to measure treatment outcomes. However, the present data also demonstrate that stable language processing patterns, as measured by eye movements, can be observed even in the presence of practice effects. Therefore, with appropriate experimental design choices, treatment-induced changes in eye movements should be detectable despite practice effects.

5. Conclusion

The results of the present study suggest that eye movements during sentence-picture matching can provide a reliable measure of sentence processing in aphasia. In unimpaired young adults, test-retest reliability was weaker, but nevertheless consistent online sentence processing patterns were observed for passive sentences, possibly reflecting thematic reanalysis processes. Thus, the present findings suggest that visual-world eyetracking may be a useful method for investigating sentence processing, particularly in individuals with aphasia, as well as for evaluating the effects of language treatment on sentence processing.

Acknowledgments

The work reported here was part of a larger, multi-site project examining the neurobiology of language recovery in people with aphasia (NIH P50-DC012283, PI: C.K. Thompson), and was also supported by NIH-DC001948 (PI: C.K. Thompson) and a Northwestern University undergraduate research grant to A.Z. Wei. The authors would like to thank the research participants and their families and caregivers, as well as Katrin Bovbjerg, Sarah Chandler, Brianne Dougherty, Mahir Mameledzija, Michaela Nerantzini, and Caitlin Radnis for assistance with data collection, and Elena Barbieri and Matthew Walenski for helpful discussions.

References

Bastiaanse, R., & Edwards, S. (2004). Word order and finiteness in Dutch and English Broca's and Wernicke's aphasia. *Brain and Language, 89*(1), 91–107.
Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.
Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences, 1191*, 133–155.
Berndt, R. S., Mitchum, C. C., & Haendiges, A. N. (1996). Comprehension of reversible sentences in "agrammatism": a meta-analysis. *Cognition, 58*(3), 289–308.
Blumenfeld, H. K., & Marian, V. (2011). Bilingualism influences inhibitory control in auditory comprehension. *Cognition, 118*(2), 245–257.
Borovsky, A., Burns, E., Elman, J. L., & Evans, J. L. (2013). Lexical activation during sentence comprehension in adolescents with history of specific language impairment. *Journal of Communnication Disorders, 46*(5–6), 413–427.
Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*(4), 417–436.
Borovsky, A., Sweeney, K., Elman, J. L., & Fernald, A. (2014). Real-time interpretation of novel events across childhood. *Journal of Memory and Language, 73*, 1–14.
Bos, L. S., Hanne, S., Wartenburger, I., & Bastiaanse, R. (2014). Losing track of time? Processing of time reference inflection in agrammatic and healthy speakers of German. *Neuropsychologia, 65*, 180–190.
Boyle, M. (2014). Test-retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research, 57*, 966–978.
Brock, J., Norbury, C., Einav, S., & Nation, K. (2008). Do individuals with autism process words in context? Evidence from language-mediated eye-movements. *Cognition, 108*(3), 896–904.
Burchert, F., & De Bleser, R. (2004). Passives in agrammatic sentence comprehension: a German study. *Aphasiology, 18*, 29–45.
Caplan, D., Michaud, J., & Hufford, R. (2013). Dissociations and associations of performance in syntactic comprehension in aphasia and their implications for the nature of aphasic deficits. *Brain and Language, 127*(1), 21–33.
Caplan, D., Waters, G., Dede, G., Michaud, J., & Reddy, A. (2007). A study of syntactic processing in aphasia I: behavioral (psycholinguistic) aspects. *Brain and Language, 101*(2), 103–150.
Caplan, D., Waters, G. S., & Hildebrandt, N. (1997). Determinants of sentence comprehension in aphasic patients in sentence-picture matching tasks. *Journal of Speech, Language, and Hearing Research, 40*(3), 542–555.
Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: evidence from aphasia. *Brain and Language, 3*(4), 572–582.
Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 1029–1040.

Cho-Reyes, S., & Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences. *Aphasiology, 26*(10), 1250–1277.
Cho, S., & Thompson, C. K. (2010). What goes wrong during passive sentence production in agrammatic aphasia: an eyetracking study. *Aphasiology, 24*(12), 1576–1592.
Choy, J. J., & Thompson, C. K. (2010). Binding in agrammatic aphasia: processing to comprehension. *Aphasiology, 24*(5), 551–579.
Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*(2), 127–137.
Dickey, M. W., Choy, J. J., & Thompson, C. K. (2007). Real-time comprehension of wh- movement in aphasia: evidence from eyetracking while listening. *Brain and Language, 100*(1), 1–22.
Dickey, M. W., & Thompson, C. K. (2009). Automatic processing of wh- and NP-movement in agrammatic aphasia: evidence from eyetracking. *Journal of Neurolinguistics, 22*(6), 563–583.
Farris-Trimble, A., & McMurray, B. (2013). Test-retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research, 56*(4), 1328–1345.
Farzin, F., Scaggs, F., Hervey, C., Berry-Kravis, E., & Hessl, D. (2011). Reliability of eye tracking and pupillometry measures in individuals with fragile X syndrome. *Journal of Autism and Developmental Disorders, 41*(11), 1515–1522.
Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology, 42*(1), 98–116.
Freed, D. B., Marshall, R. C., & Chuhlantseff, E. A. (1996). Picture naming variability: a methodological consideration of inconsistent naming responses in fluent and nonfluent aphasia. *Clinical Aphasiology, 24*, 193–206.
Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *Various coefficients of interrater reliability or agreement*. Retrieved from http://cran.r-project.org/web/packages/irr/irr.pdf.
Grodzinsky, Y. (1986). Language deficits and the theory of syntax. *Brain and Language, 27*(1), 135–159.
Grodzinsky, Y., Piñango, M. M., Zurif, E., & Drai, D. (1999). The critical role of group studies in neuropsychology: comprehension regularities in Broca's aphasia. *Brain and Language, 67*(2), 134–147.
Guo, C. C., Kurth, F., Zhou, J., Mayer, E. A., Eickhoff, S. B., Kramer, J. H., et al. (2012). One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. *Neuroimage, 61*(4), 1471–1483.
Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2015). Sentence comprehension and morphological cues in aphasia: what eye-tracking reveals about integration and prediction. *Journal of Neurolinguistics, 34*, 83–111.
Hanne, S., Sekerina, I. A., Vasishth, S., Burchert, F., & De Bleser, R. (2011). Chance in agrammatic sentence comprehension: what does it really mean? Evidence from eye movements of German agrammatic aphasic patients. *Aphasiology, 25*(2), 221–244.
Hirotani, M., Makuuchi, M., Ruschemeyer, S. A., & Friederici, A. D. (2011). Who was the agent? The neural correlates of reanalysis processes during sentence comprehension. *Human Brain Mapping, 32*(11), 1775–1787.
Hochstadt, J. (2009). Set-shifting and the on-line processing of relative clauses in Parkinson's disease: results from a novel eye-tracking method. *Cortex, 45*(8), 991–1011.
Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologica, 137*(2), 151–171.
Kamide, Y., Scheepers, C., & Altmann, G. T. (2003). Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research, 32*(1), 37–55.
Kertesz, A. (2006). *Western Aphasia Battery-Revised (WAB-R)*. San Antonio, TX: Pearson.
Kim, E. S., & Lemke, S. F. (2016). Behavioural and eye-movement outcomes in response to text-based reading treatment for acquired alexia. *Neuropsychological Rehabilitation, 26*(1), 60–86.
Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition, 95*(1), 95–127.
Kuznetsova, A., Brockhoff, P. B., & Christensen, P. H. B. (2015). *lmerTest: Tests in linear mixed effects models*. https://cran.r-project.org/web/packages/lmerTest/index.html.
Laurinavichyute, A. K., Ulicheva, A., Ivanova, M. V., Kuptsova, S. V., & Dragoy, O. (2014). Processing lexical ambiguity in sentential context: Eye-tracking data from brain-damaged and non-brain-damaged individuals. *Neuropsychologia, 64*, 360–373.
Lawrence, M. A. (2013). *ez: Easy analysis and visualization of factorial experiments*. Retrieved from https://cran.r-project.org/web/packages/ez/index.html.
Lee, C. I., Mirman, D., & Buxbaum, L. J. (2014). Abnormal dynamics of activation of object use information in apraxia: evidence from eyetracking. *Neuropsychologia, 59*, 13–26.
Lee, J., & Thompson, C. K. (2011a). Real-time production of arguments and adjuncts in normal and agrammatic speakers. *Language and Cognitive Processes, 26*(8), 985–1021.
Lee, J., & Thompson, C. K. (2011b). Real-time production of unergative and unaccusative sentences in normal and agrammatic speakers: an eyetracking study. *Aphasiology, 25*(6–7), 813–825.
Luck, S. J. (2014). *An introduction to the event-related potential technique, second edition*. Cambridge, MA: MIT Press.
Mack, J. E., Ji, W., & Thompson, C. K. (2013). Effects of verb meaning on lexical integration in agrammatic aphasia: evidence from eyetracking. *Journal of Neurolinguistics, 26*(6), 619–636.
Mack, J. E., Meltzer-Asscher, A., Barbieri, E., & Thompson, C. K. (2013). Neural correlates of processing passive sentences. *Brain Sciences, 3*(3), 1198–1214.
Malyutina, S., & den Ouden, D. B. (2015). What is it that lingers? Garden-path (mis)interpretations in younger and older adults. *Quarterly Journal of Experimental Psychology*, 1–27.
Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake - But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 843–847.
Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science, 11*(3), F9–F16.
McMurray, B., Munson, C., & Tomblin, J. B. (2014). Individual differences in language ability are related to variation in word recognition, not speech perception: evidence from eye movements. *Journal of Speech, Language, and Hearing Research, 57*(4), 1344–1362.
McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: implications for SLI. *Cognitive Psychology, 60*(1), 1–39.
McNeil, M. R., Pratt, S. R., Szuminsky, N., Sung, J. E., Fossett, T. R. D., Fassbinder, W., et al. (2015). Reliability and validity of the Computerized Revised Token Test: comparison of reading and listening versions in persons with and without aphasia. *Journal of Speech, Language, and Hearing Research, 58*(2), 311–324.
Meyer, A. M., Mack, J. E., & Thompson, C. K. (2012). Tracking passive sentence comprehension in agrammatic aphasia. *Journal of Neurolinguistics, 25*(1), 31–43.
Mirman, D., & Graziano, K. M. (2012). Damage to temporo-parietal cortex decreases incidental activation of thematic relations during spoken word comprehension. *Neuropsychologia, 50*, 1990–1997.
Mirman, D., Irwin, J. R., & Stephen, D. G. (2012). Eye movement dynamics and cognitive self-organization in typical and atypical development. *Cognitive Neurodynamics, 6*(1), 61–73.
Mirman, D., Yee, E., Blumstein, S. E., & Magnuson, J. S. (2011). Theories of spoken word recognition deficits in aphasia: evidence from eye-tracking and computational modeling. *Brain and Language, 117*, 53–68.

Nation, K., Marshall, C. M., & Altmann, G. T. (2003). Investigating individual differences in children's real-time sentence comprehension using language-mediated eye movements. *Journal of Experimental Child Psychology, 86*(4), 314–329.
Patil, U., Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2015). A computational evaluation of sentence processing deficits in aphasia. *Cognitive Science, 40*, 5–50.
Pickering, M. J., McLean, J. F., & Branigan, H. P. (2013). Persistent structural priming and frequency effects during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 890–897.
R Core Team. (2015). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org/.
Raffray, C. N., & Pickering, M. J. (2010). How do people construct logical form during language comprehension? *Psychological Science, 21*(8), 1090–1097.
Sheppard, S., Walenski, M., Love, T., & Shapiro, L. P. (2015). The auditory comprehension of wh-questions in aphasia: support for the intervener hypothesis. *Journal of Speech, Language, and Hearing Research, 58*, 781–797.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.
Thompson, C. K. (2011). *Northwestern Assessment of Verbs and Sentences (NAVS)*. Evanston, IL.
Thompson, C. K., Cho, S., Hsu, C. J., Wieneke, C., Rademaker, A., Weitner, B. B., et al., … (2012). Dissociations between fluency and agrammatism in primary progressive aphasia. *Aphasiology, 26*(1), 20–43.
Thompson, C. K., & Choy, J. J. (2009). Pronominal resolution and gap filling in agrammatic aphasia: evidence from eye movements. *Journal of Psycholinguistic Research, 38*(3), 255–283.
Thompson, C. K., Meltzer-Asscher, A., Cho, S., Lee, J., Wieneke, C., Weintraub, S., et al. (2013). Syntactic and morphosyntactic processing in stroke-induced and primary progressive aphasia. *Behavioural Neurology, 26*(1–2), 35–54.
Thompson, C. K., & Weintraub, S. (2014). *Northwestern Naming Battery (NNB)*. Evanston, IL.
Venker, C. E., Eernisse, E. R., Saffran, J. R., & Ellis Weismer, S. (2013). Individual differences in the real-time comprehension of children with ASD. *Autism Research, 6*(5), 417–432.
Villard, S., & Kiran, S. (2015). Between-session intra-individual variability in sustained, selective, and integrational non-linguistic attention in aphasia. *Neuropsychologia, 66*, 204–212.
Yee, E., Blumstein, S. E., & Sedivy, J. C. (2008). Lexical-semantic activation in Broca's aphasia: evidence from eye movements. *Journal of Cognitive Neuroscience, 20*(4), 592–612.
Zanto, T. P., Pa, J., & Gazzaley, A. (2014). Reliability measures of functional magnetic resonance imaging in a longitudinal evaluation of mild cognitive impairment. *Neuroimage, 84*, 443–452.