Supplementary Table S1 provides a list of verbs trained during the treatment blocks, in each language.

Supplementary Appendix S1, Appendix S2, Table S2, Figures S1, S2, S3 and S4 provide a description of our methods and results for calculating whether learning occurred during each treatment block for both participants.

Supplementary Table S3 provides a detailed summary of the Aphasia Battery subtests used in this study.

Supplementary Table S1. *List of trained verbs.*

ENGLISH				
Verbs in the sentence construction subtest of the testing battery:				
(1) Weigh	(2) Wash	(3) Catch	(4) Open	(5) Kick
(6) Plant	(7) Guard	(8) Stop	(9) Pour	(10) Carry
Verbs not in the testing battery in any subtest:				
(1) Bake	(2) Lose	(3) Lend	(4) Leave	(5) Watch
(6) Find	(7) Fight	(8) Examine	(9) Shout	(10) Hug
Extra verbs switched into the second treatment block:				
(1) Accept	(2) Understand	(3) Drop		
HEBREW				
Verbs in the sentence construction subtest of the testing battery:				
(1) שוקל	(2) שוטף	(3) תופס	(4) פותח	(5) בועט
(6) שותל	(7) שומר	(8) עוצר	(9) מוזג	(10) סוחב
Verbs not in the testing battery in any subtest:				
(1) אופה	(2) מאבד	(3) משאיל	(4) עוזב	(5) צופה
(6) מוצא	(7) נאבק	(8) בוחן	(9) צועק	(10) מחבק
Extra verbs switched into the second treatment block:				
(1) מקבל	(2) מבין	(3) מפיל		

Note. The verbs are numbered to allow for comparison across English and Hebrew.

Supplementary Appendix S1. *Methodology for calculating improvement during the treatment blocks.*

To demonstrate whether the participants improved in independently producing agents and patients when given a verb and/or producing relevant SVO sentences during the VNeST protocol, we looked at two measures directly from the treatment blocks. First, we tallied the number of times agents or patients were retrieved independently, and/or the number of times they were retrieved after either a minimal or maximal cue, during the early stage of each verb cycle. Since the participants were asked to produce four sentences per verb, with two thematic roles per verb (one agent and one patient), this measure was calculated as an average per session (max = 8 thematic roles per verb). Second, we tallied the number of relevant SVO sentences produced independently at the final stage of each verb cycle (maximum of 4 per verb). This is a novel way of measuring direct improvement during treatment blocks. Previous studies with VNeST have used other tasks to try to capture direct treatment effects, such as using trained and untrained sentence probes, or sentence probes relative to control sentence probes during treatment blocks (e.g., Edmonds et al., 2009; Edmonds & Babb, 2011; Edmonds et al., 2014). However, these methods both considerably increased participant frustration (observed in previous studies by the third author, and in our pilot participant for the current study) and do not directly measure treatment effects.

Non-parametric correlations were calculated to examine change in the treated language. We analysed the correlation between (1) the average number of thematic role fillers (i.e., agents and/or patients) retrieved independently and treatment session number within a treatment block, (2) the average number of thematic role fillers retrieved after a minimal cue and treatment session number within a treatment block, and (3) the average number of thematic role fillers

retrieved after a maximal cue and treatment session number within a treatment block. We expected that positive correlations would be observed for independent retrieval, so that as treatment sessions progressed, more independent retrieval would occur due to the treatment. We also expected that negative correlations would be observed for retrieval after a minimal or maximal cue as treatment sessions progressed because, based on the VNeST protocol, as independent retrieval improves, less cueing is necessary. Non-parametric correlations were also calculated between the average number of relevant SVO sentences produced independently and treatment session number within a treatment block. We expected these correlations to be positive, such that as treatment sessions progressed, more relevant SVO sentences production would occur, due to the treatment. We used the Benjamini-Hochberg approach to multiple comparisons because it controls the family-wise error rate without being overly conservative (Benjamini & Hochberg, 1995).

Supplementary Appendix S2. *Results*.

For both EH03 and EH04 correlations trended in the directions predicted: positive for correlations between retrieval of agents and patients independently and treatment session number, positive between the independent retrieval of relevant SVO sentences per verb and treatment session number, negative between retrieval of agents and patients after a minimal cue and treatment session number, and negative between retrieval of agents and patients after a maximal cue and treatment session number. Some correlations were observed to be significant, for both EH03 and EH04:

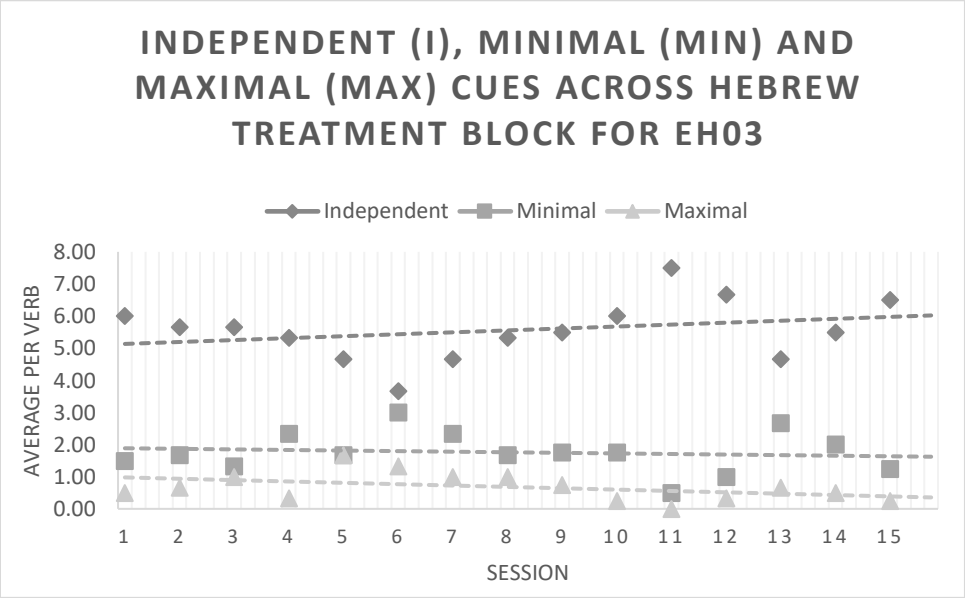
For EH03 in Hebrew (his first treatment block), the correlation between the average number of relevant SVO sentences produced independently per verb and treatment session number was positive and significant, as predicted. In English (his second treatment block) correlations were significant between session and independent retrieval (positive, as predicted). For EH04 in English (his first treatment block) for the measure of retrieval of agents and patients, the correlations between session and independent retrieval and between session and retrieval after a maximal cue were significant (as predicted, positive for independent retrieval and negative for retrieval after a maximal cue). In Hebrew (his second treatment block) no measure was observed to be significant. See Table ST3 for a summary of the correlation calculations, and Figures SF1, SF2, SF3 and SF4 that show the trajectory of change over time on agent and patient retrieval.

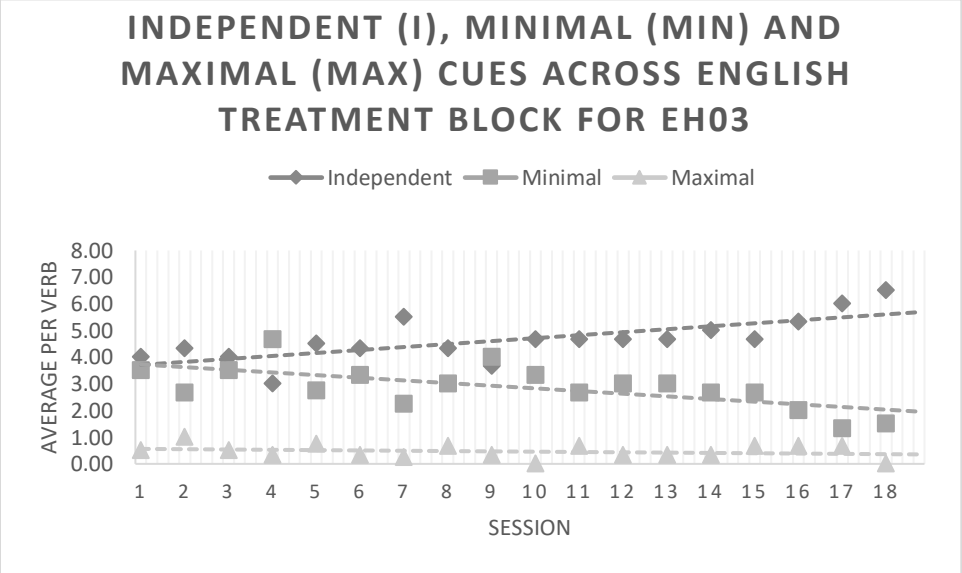
Supplementary Table S2. *Changes to retrieval of agents and patients in each treatment block for EH03 and EH04.*

	EH03		**EH04**	
	Treatment block 1 (Hebrew)	Treatment block 2 (English)	Treatment block 1 (English)	Treatment block 2 (Hebrew)
Independent retrieval of agents and patients across sessions	Not significant	**Positive and significant**	**Positive and significant**	Approached positive significance
	r(14) = .286, *p* = .301	**r(17) = .725, *p* < .001**	**r(15) = .603, *p* = .013**	r(13) = .524, *p* = .054
Retrieval of agents and patients after a minimal cue across sessions	Not significant	Approached negative significance	Not significant	Approached negative significance
	r(14) = -.124, *p* = .659	r(17) = -.651, *p* = .003	r(15) = -.270, *p* = .312	r(13) = -.539, *p* = .047
Retrieval of agents and patients after a maximal cue across sessions	Not significant	Rarely needed – not significant	**Negative and significant**	Rarely needed – not significant
	r(14) = -.419, *p* = .120	r(17) = -.239, *p* = .339	**r(15) = -.810, *p* < .001**	r(13) = -.331, *p* = .247
Average number of relevant SVO sentences produced independently per verb across sessions	**Positive and significant**	Not significant	Approached positive significance	Not significant (close to ceiling)
	r(14) = .709, *p* =.003	r(18) = .419, *p* =.083	r(15) = .512, *p* = .043	r(13) = .076, *p* = .796

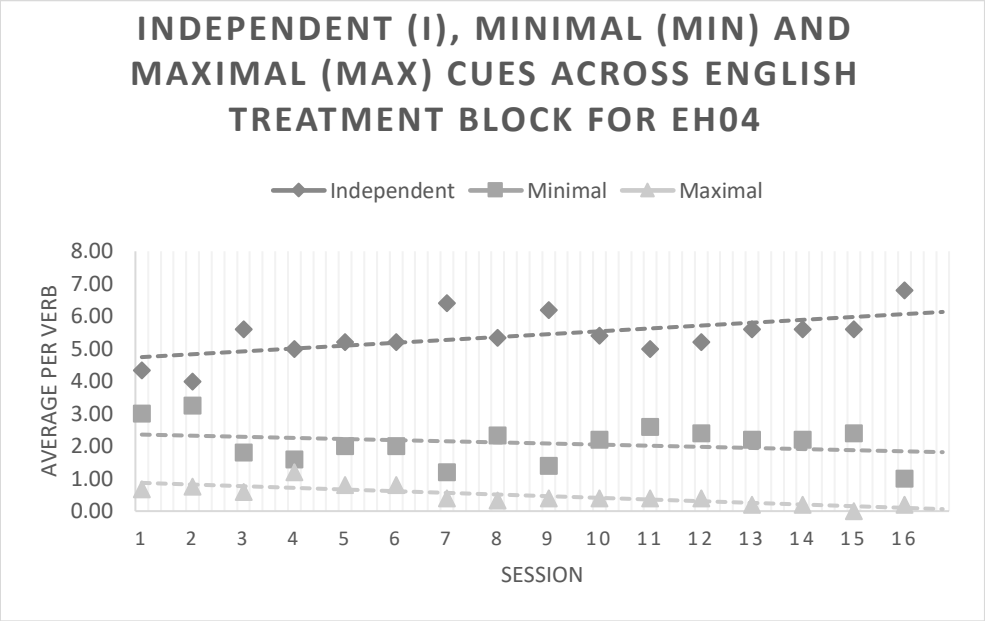
Note. r = correlation between retrieval and treatment session number; bold indicates statistical significance at the .05 level, corrected for multiple comparisons using the Benjamini-Hochberg approach.

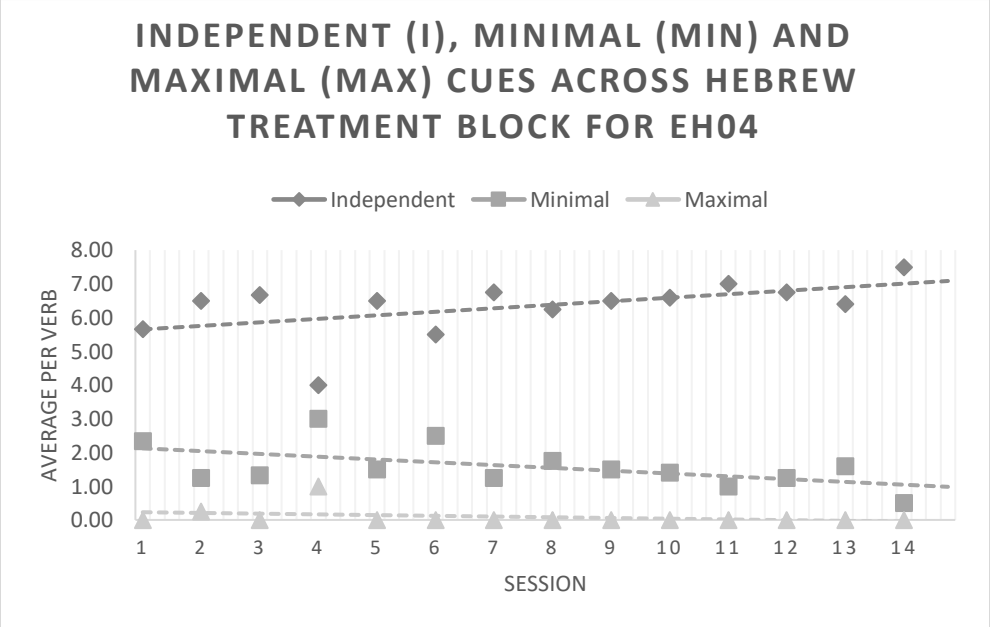
Supplementary Figure S1. *Independent, minimal, and maximal cues across the Hebrew and English treatment blocks for EH03.*



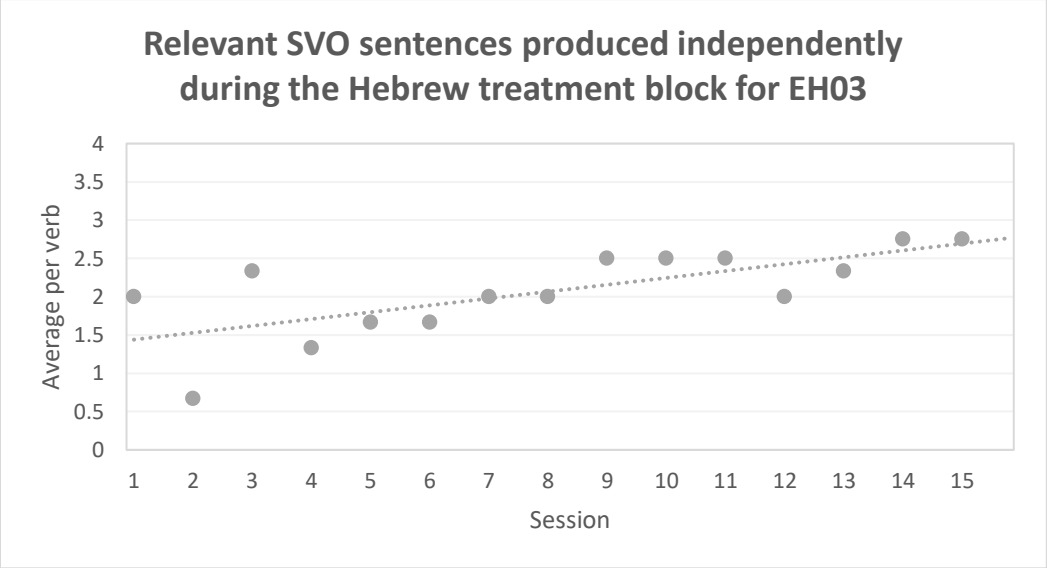


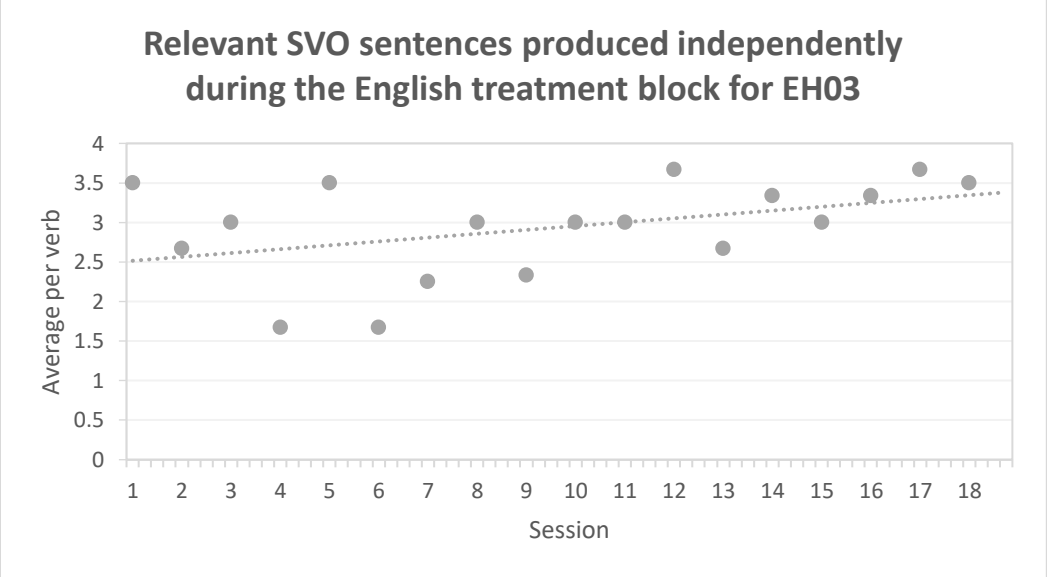
Supplementary Figure S2. *Independent, minimal, and maximal cues across the English and Hebrew treatment blocks for EH04.*



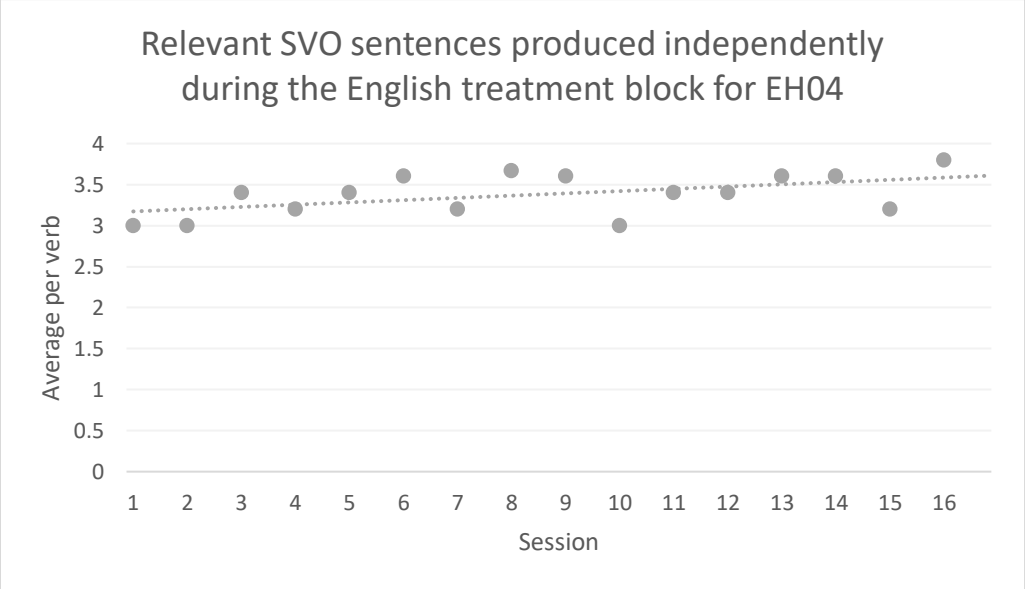


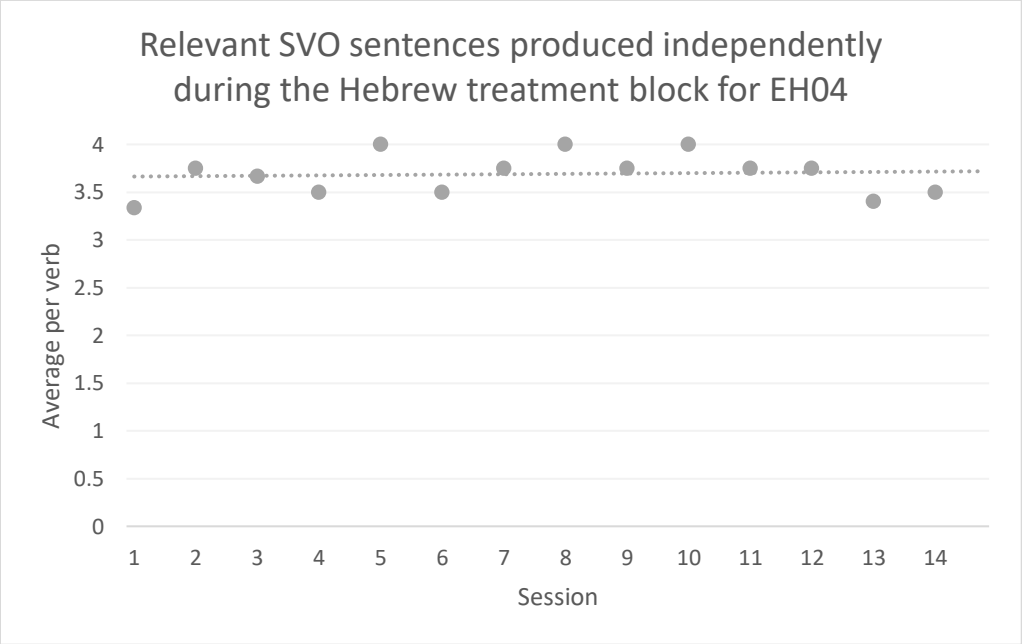
Supplementary Figure S3. *Independent relevant SVO sentence production across the Hebrew and English treatment blocks for EH03.*





Supplementary Figure S4. *Independent relevant SVO sentence production across the English and Hebrew treatment blocks for EH04.*





Supplementary Table S3. *The subtests of the Revised English-Hebrew Aphasia Battery (REHAB).*

Production tasks:
1. Picture-based action naming (n = 45). The participants were asked to name each action. Pictures were partially based on a subset of actions from the Action Naming Test which is a subtest of the Verb and Sentence Test (Bastiaanse et al., 2002). Other pictures of actions were also selected - all 45 actions were non-cognates in English and Hebrew. The whole set of pictures was piloted on two healthy speakers of English and two healthy speakers of Hebrew, to confirm that the pictures elicited the intended verbs.
2. Picture-based object naming (n = 45). The participants were asked to name each object. Pictures were partially based on a subset of objects from the Multilingual Naming Test (Gollan et al., 2012). Other pictures of objects were also selected - all 45 objects were non-cognates in English and Hebrew. The whole set of pictures was piloted on two healthy speakers of English and two healthy speakers of Hebrew, to confirm that the pictures elicited the intended nouns.
3. Picture-based sentence construction (n=54). Pictures were partially based on a subset of pictures from the Object and Action Naming Battery (Druks & Masterson, 2000). Other sentence-eliciting pictures were also selected - all 54 pictures included actions that were non-cognates in English and Hebrew. The whole set of pictures was piloted on two healthy speakers of English and two healthy speakers of Hebrew, to confirm that the pictures elicited 12 1-argument sentences, 30 pictures elicited 2- argument sentences, and 12 pictures elicited 3-argument sentences.
4. Discourse (n = 15) included six single picture descriptions, three 6-picture story sequence, three requests for personal information (e.g., “tell me about a family

vacation" and three requests for procedural information (e.g., how do you make an omelette?"), as recommended by Nicholas & Brookshire (1993).

Three different sets of stimuli were developed for oral production of discourse, and included picture descriptions of the "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983) and the "picnic" picture from the WAB-R (Kertesz, 2006), as well as the 6-picture story sequence from the Bilingual Aphasia Test (Paradis, 2011). These pictures and story sequences were piloted together with new picture description and story sequence stimuli specifically drawn or selected to elicit discourse with varied and rich vocabulary. The pilot participants were 12 healthy speakers (f = 7, m = 5): five native speakers of Hebrew (providing data in Hebrew only) and seven native speakers of English (four British English, three American English, providing data in English only). They were aged 57-76 years (mean 68.3 years), had 11-25 years of formal education (mean 16.5), and were self-reported to have middle to high socioeconomic status. The pilot data indicated that the three sets of discourse stimuli were comparable to each other and across languages.

Comprehension tasks:

(The participants were asked to point to the picture corresponding to the auditory stimulus, out of four options.)

1. Auditory comprehension of nouns (based on a subset of stimuli from the Peabody Picture Vocabulary Test; Dunn & Dunn, 2007)
2. Auditory comprehension of verbs (based on a subset of stimuli from the Verb and Sentence Test; Bastiaanse, Edwards, & Rispens, 2002). Distractors were a semantically related verb and two nouns semantically related to each verb (the target and the distractor).

References:

Bastiaanse, R., Edwards, S., & Rispens, J. (2002). *Verb and Sentence test (VAST)*. Suffolk: Thames Valley Test Company.

Druks, J., & Masterson, J. (2000). *An Object and Action Naming Battery*. Hove, UK: Psychology Press.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test (4th ed.).* San Antonio: Pearson.

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *15*(03), 594–615. https://doi.org/10.1017/S1366728911000332

Goodglass, H., & Kaplan, E. (1983). *Boston Diagnostic Aphasia Examination, Vol. 2nd ed.* Media, PA: Williams and Wilkins.

Kertesz, A. (2006). *Western Aphasia Battery—Revised*. Austin, TX: Pro-Ed.

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech, Language, and Hearing Research*, *36*(2), 338–350.

Paradis, M. (2011). Principles underlying the Bilingual Aphasia Test (BAT) and its uses. *Clinical Linguistics & Phonetics*, *25*(6–7), 427–443. https://doi.org/10.3109/02699206.2011.560326