Verbal recall captures neural encoding of a drug-themed movie in heroin use disorder

Sarah G. King[1], Greg Kronberg[1], Natalie McClain[2], Zihan Zhang[1,3], Ahmet O. Ceceli[1], Yuefeng Huang[1], Nelly Alia-Klein[1], Rita Z. Goldstein[1,2]

[1] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[2] Department of Neuroscience, Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[3] Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Corresponding author: Rita Z. Goldstein (rita.goldstein@mssm.edu)

Abstract

Humans use language to express complex thoughts and internal states, offering a window into dynamic neurocognitive processes underlying psychiatric conditions. For people with substance use disorders, intense emotional and physiological responses to a drug-related context can precipitate craving and drug seeking, yet sensitive and accessible behavioral markers of the underlying neural dynamics remain elusive. Here, we examined verbal recall of a drug-themed movie in inpatients with heroin use disorder (HUD) and healthy controls to characterize linguistic representations of the processing of a naturalistic stimulus. The HUD participants showed greater intersubject semantic variability and more drug-biased and self-referential, but less socially directed, language; except for drug bias, these differences diminished with three months in treatment. Semantic variability paralleled functional variabilities in dorsal attention and visual networks during movie viewing, linking motivated attentional processing to later recall. Semantic-neural correspondences were consistent between timepoints in HUD participants in the superior parietal lobule, suggesting partial stability in the pattern of attention network effects. These findings support natural language as a potential scalable marker of neural encoding for clinical monitoring in addiction.

Introduction

Language is a uniquely human behavior that can provide ecologically valid insights into subjective experiences that are difficult to capture in laboratory settings. Recent advances in natural language processing, particularly with transformer-based large language models, enable the extraction of latent semantic patterns from spontaneous speech and narrative recall [1]. These tools offer promising approaches for facilitating enhanced phenotyping, outcome prediction, and digital monitoring of psychiatric conditions [2–7]. By engaging distributed neural systems [8,9] and recruiting functions like attention, self-regulation, and perspective-taking [10], natural language may be especially sensitive to variations in cognitive states, providing clinically meaningful markers of disease [11–13]. Despite this potential, language and naturalistic behaviors remain underutilized for studying the underlying mechanisms of drug addiction [14], a condition for which subjective experiences are foundational to both neurocognitive underpinnings and clinical trajectories.

Critically, the opioid epidemic accounted for approximately 90% of the estimated 105,007 drug overdose-related deaths in the United States in 2023 [15]. Only a minority of individuals with opioid use disorder typically receive any form of treatment [16], and relapse rates remain high even after sustained abstinence [17]. A core neurocognitive impairment in addiction is enhanced salience attribution to drug-related stimuli at the expense of alternative reinforcers; accompanied by impaired behavioral control over drug use, this presents a significant barrier to recovery [18–20]. Experimental cue reactivity paradigms consistently demonstrate drug-biased activations in large-scale networks associated with reward, salience, executive, and self-directed functions, while these same systems are downregulated in nondrug-related social/emotional, inhibitory control, and decision-making tasks [21]. Such dynamics contribute to cue-induced

relapse, whereby real-world drug-related contexts provoke craving and diminish control over further use [22]. In our prior work [23], using a naturalistic movie stimulus, we identified drug-biased intersubject response synchronizations between individuals with heroin use disorder (HUD) in regions associated with reward, salience, and inhibitory control networks. Synchrony in the orbitofrontal cortex decreased with treatment and tracked reductions in self-reported drug cue-induced cravings during treatment, highlighting the value of naturalistic paradigms for capturing real-world neurocognitive dynamics of drug cue processing.

Here, we extend this framework by focusing on verbal recall of the same stimulus, which constituted part of the movie *Trainspotting* (1996) and included both explicit and implicit drug-related themes embedded within relevant social and emotional contexts. After watching the movie excerpt, participants verbally recalled the narrative and reflected on their subjective experience (e.g., emotional reactions and feelings of personal relevance) of the stimulus. Using a transformer-based embeddings model, we tested the discrimination between the HUD and healthy control (CTL) groups, and between inpatient treatment time points, in latent contextual features of participants' recall. We then quantified specific linguistic markers reflecting drug-biased attention (drug-related words) and self- versus socially directed processing (pronoun use). Finally, we applied representational similarity analysis to test whether latent semantic features in recall corresponded to fMRI response patterns during movie viewing, evaluating language as a potential behavioral correlate of drug context encoding in the brain.

Results

Participant Characteristics

Participant groups did not differ significantly in age, gender, race, or days between sessions, but did differ in education, verbal and nonverbal IQ (CTL>HUD), and in depression and anxiety symptoms and alcohol dependence (HUD>CTL). Since all but one CTL participant and all but two HUD participants were/were not regularly using nicotine, respectively, no statistical comparison was performed for nicotine use. Statistical tests between the groups for these variables and descriptive statistics for other drug use/addiction severity measures in the HUD group are presented in Table 1 (see Table S4 for the baseline replication sample, which included participants who only had Session 1 data, described in Methods). Correlations were inspected between the language dependent variables (e.g., semantic similarity and term frequency measures) and participant variables that were not relevant to our analyses, but which differed significantly between the groups (e.g., education, verbal and nonverbal IQ, and alcohol dependence); none of these correlations were significant (all $p>.05$), hence these variables were not considered as covariates.

Table 1 – Baseline demographic, neuropsychological, and drug use/addiction measures

	CTL, n=31	HUD, n=40	Test statistic	p-value
Demographics				
Age, mean (SD) years	40.0 (11.0)	42.1 (11.0)	W=543	.375
Gender, # men/women	20/11	30/10	χ^2=0.92	.337
Race, # Black/white/multiracial or other	8/15/8	7/24/9	χ^2=1.08	.583
Education, mean (SD) years	16.3 (3.0)	12.2 (2.5)	W=1048	<.001
Neuropsychological tests				
WRAT – mean (SD) Reading standard score	109.6 (6.5)	95.6 (11.5)	W=1062	<.001
WASI – mean (SD) Matrix Reasoning scaled score	11.8 (2.6)	9.9 (3.0)	W=853	.007
Depression symptoms, mean (SD) BDI score[a]	4.6 (9.6)	12.8 (11.1)	W=239	<.001
Anxiety symptoms, mean (SD) BAI score[a]	3.3 (6.7)	9.3 (8.6)	W=281	<.001
Drug use and addiction variables				
Regular nicotine use, # current/past/never[b]	1/4/26	38/2/0	-	-
Alcohol dependence, mean (SD) SMAST score	0.3 (0.9)	2.1 (2.5)	W=205	<.001
Heroin use disorder				
Age of onset, mean (SD) years	-	25.8 (7.2)	-	-
Regular use, mean (SD) years	-	10.8 (8.1)	-	-
Current abstinence, mean (SD) days[a]	-	235.5 (260.7)	-	-
Past 30-day use, mean (SD) days[a]	-	0.2 (0.7)	-	-
Craving symptoms, mean (SD) HCQ score	-	37.6 (13.9)	-	-
Withdrawal symptoms, mean (SD) SOWS score	-	2.5 (2.9)	-	-
Severity of dependence, mean (SD) SDS score	-	10.7 (3.9)	-	-
Session interval, mean (SD) days	101.6 (69.9)	95.2 (20.5)	W=487	.124

Notes: Table presents mean and standard deviation (SD) or quantities for baseline sample characteristics. Test statistics were derived from two-sample Wilcoxon signed-rank test (W) for continuous variables and chi-squared test (χ^2) for categorical variables. [a]Baseline depression, anxiety, and stress scores and abstinence duration/recent heroin use are presented here; session differences for these measures in individuals with HUD are presented in Table 2. [b]Since only one CTL participant reported current nicotine use and only two HUD participants reported non-current nicotine use, no statistical test was performed on this measure. Missing data: 1 CTL datapoint for education; 1 datapoint for regular heroin use.

Changes in Craving and Severity Measures with Treatment

Comparisons between sessions in the HUD group showed no significant changes over time in depression and anxiety symptoms. As expected for this inpatient population, the mean abstinence duration was significantly longer at Session 2, and there was no significant difference in self-reported past-month heroin use, which had a mean of <1 day at both sessions. Also as expected, there were significant reductions at Session 2 in the pre-movie (baseline) and post-movie (task-induced) craving and the mean scene-induced craving ratings. No significant session

difference was observed in the magnitude of the pre-to-post movie change in craving. Statistics for these session comparisons are presented in Table 2.

Table 2 – Repeated neuropsychological and heroin use/craving outcome measures in participants with heroin use disorder

	Session 1	Session 2	Test statistic	p-value
Neuropsychological tests				
Depression symptoms, mean (SD) BDI score[a]	12.8 (11.1)	11.4 (9.8)	V=414	.206
Anxiety symptoms, mean (SD) BAI score[a]	9.3 (8.6)	10.6 (9.4)	V=274	.694
Heroin use and craving variables				
Current abstinence, mean (SD) days	235.5 (260.7)	310.1 (281.8)	V=86	<.001
Past 30-day use, mean (SD) days	0.2 (0.7)	0.7 (2.5)	V=11	.357
Pre-movie/baseline craving, mean (SD) rating[a]	2.44 (2.16)	1.44 (1.82)	V=303	.006
Post-movie/task-induced craving, mean (SD) rating[a]	3.00 (2.47)	2.03 (2.43)	V=289	.016
Movie craving change, mean (SD) rating difference	0.56 (1.83)	0.59 (1.46)	V=151	.697
Scene-induced craving, mean (SD) rating[a]	1.19 (1.20)	0.75 (0.99)	V=408	.002

Notes: Table presents mean and standard deviation (SD) for repeated measures in the HUD group. Test statistics were derived from paired Wilcoxon signed-rank test (V). [a]Selected outcome measures were assessed for correlations with test variables of interest. Missing data: 1 datapoint for BDI and BAI; 1 datapoint for each of the movie craving measures.

Decoding Group and Session from Contextual Embeddings

Classification analysis was performed on the first 25 principal components of 768-dimensional contextual embeddings vectors to test the sample discrimination by group and session labels (Figure 1; see Figure S2 for the baseline replication sample). A logistic regression classifier performed significantly better than chance ($p<.001$) in classifying all 142 samples by both group and session (4 classes), yielding a true accuracy score of 0.49 with a mean null accuracy score of 0.26. In follow-up analyses, classification by group (71 samples/2 classes) was significantly better than chance in both Session 1 (true accuracy: 0.86, mean null accuracy: 0.51, $p<.001$) and Session 2 (true accuracy: 0.81, mean null accuracy: 0.50, $p<.001$). Classification by session (2 classes) was significantly better than chance for the 80 samples from the HUD group

(true accuracy: 0.64, mean null accuracy: 0.50, p=.013) but not the 62 samples from the CTL group (true accuracy: 0.47, mean null accuracy: 0.49, p=.657).

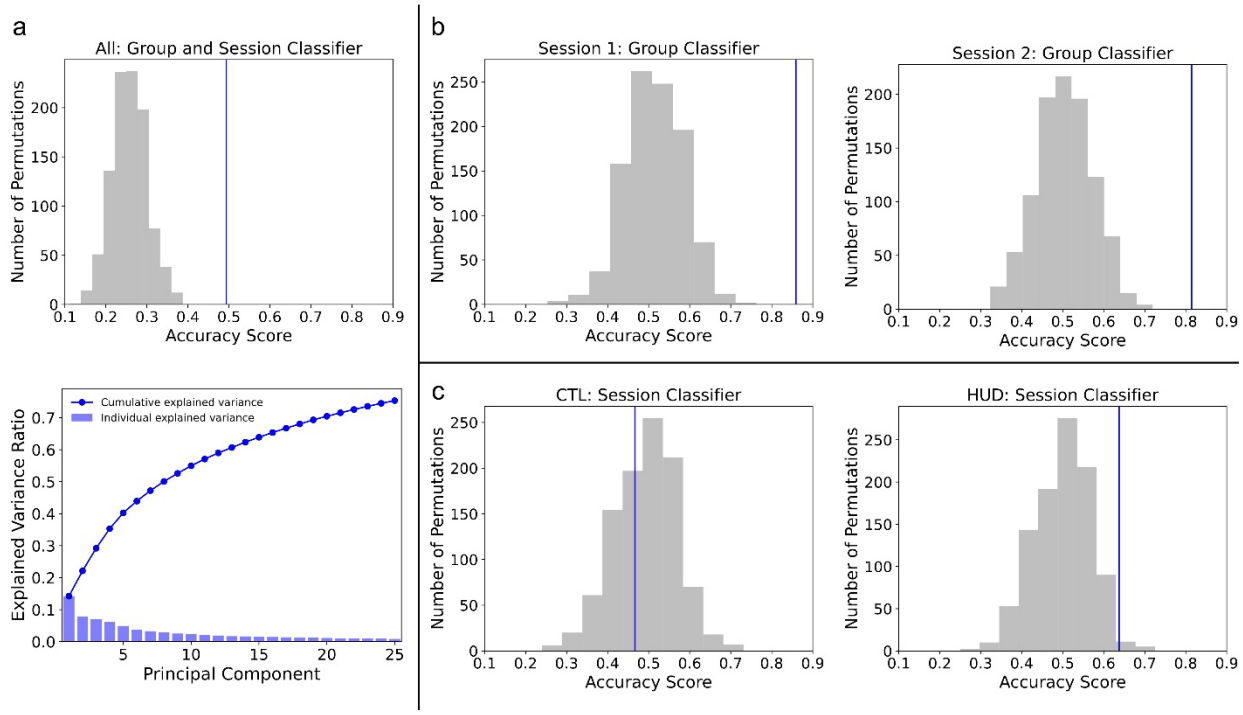

Figure 1. Classifier performance on the first 25 principal components of the embeddings, which explained 75.4% of the total variance (variance plot). In the classifier plots, the blue line indicates the true accuracy score, and the gray histogram represents the null accuracy distribution derived from 1,000 permutations of shuffled labels. Classification accuracy was significantly better than chance in decoding a: both the group and session labels from all samples (0.49), b: group labels from Session 1 (0.86) and Session 2 (0.81) samples, and c: session labels from HUD (0.64), but not CTL (0.47), samples.

Group and Session Comparisons in Semantic Similarity

Semantic similarity of the embeddings was assessed within (CTL-CTL, HUD-HUD) and between (HUD-CTL) the groups (Figure 2b; see Figure S3 for the baseline replication sample),

and within subjects (i.e., between the two sessions for the same subject) (Figure 2c). A mixed ANOVA for Comparison (CTL-CTL, HUD-HUD, HUD-CTL) and Session (1, 2) revealed significant main effects of both Comparison ($F_{1,78}$=8.13, p=.006) and Session ($F_{1,78}$=12.79, p<.001), with a significant 2-way interaction effect ($F_{2,78}$=5.28, p=.025). Follow-up 1-way ANOVAs showed a significant main effect of Comparison in Session 1 ($F_{2,108}$=8.21, p<.001), with the Tukey HSD post hoc test showing CTL-CTL>HUD-HUD (p<.001), CTL-CTL>HUD-CTL (p=.003), and HUD-HUD=HUD-CTL (p=.876); there was no significant effect at Session 2 ($F_{2,108}$=1.70, p=.189). Paired t-tests for the effect of Session showed Session 1=2 for CTL-CTL (t_{30}=0.09, p=.927) and Session 2>1 for HUD-HUD ($t_{39=}$-4.19, p<.001) and HUD-CTL (t_{39}=-2.74, p=.009). For the within-subject/between-sessions cosine similarity, an independent groups t-test showed CTL>HUD (t_{69}=3.82, p<.001).

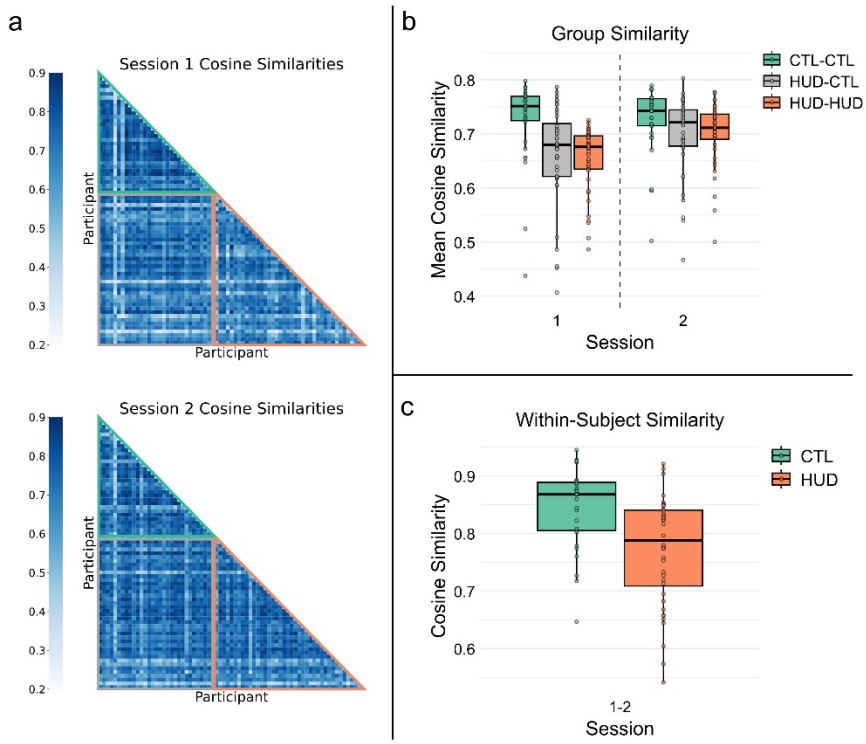

Figure 2. Semantic similarity derived from the embeddings vectors of the speech samples. a: Lower triangle of the pairwise similarity matrices from Session 1 and Session 2. Rows and

columns represent the 71 participants, sorted by group (green upper border: CTL; orange lower border: HUD), with pairwise cosine similarity values indicated by the scale bar. b. Mean pairwise CTL-CTL similarity was significantly higher than HUD-CTL and HUD-HUD similarities at Session 1, with no significant differences at Session 2, and both HUD-CTL and HUD-HUD similarities increased between the sessions. c: CTL within-subject similarity was significantly higher than HUD, indicating greater change in HUD participants' speech between the two sessions. In the boxplots, the center line indicates the median, the box indicates the interquartile range, and whiskers extend to the most extreme points within 1.5x the interquartile range above and below the box.

Group and Session Comparisons on Drug Bias and Self vs. Other Reference

To better understand the observed patterns of group and session effects in the embeddings, we examined semantic features that were hypothesized to occur more frequently in the HUD group's recall. These included drug-biased speech and self-reference vs. socially-directed speech, which were approximated as the frequencies of drug terms and first-person (FP) vs. third-person (TP) pronouns, respectively, in the sample (Figure 3; see Figure S4 for the baseline replication sample). For drug term frequencies, a mixed ANOVA examining Group (CTL, HUD) and Session (1, 2) revealed a significant main effect of Group (HUD>CTL, $F_{1,69}$=5.84, p =.018), and no other main or interaction effects ($F \leq 0.44$, $p \geq .510$).

The mixed ANOVA examining pronoun term frequencies by Group (CTL, HUD), Session (1, 2), and Type (FP, TP) revealed a significant main effect of Type ($F_{1,69}$=25.86, p<.001), a significant 2-way interaction effect of Session × Type ($F_{1,69}$=13.65, p<.001), and a significant 3-way interaction effect of Group × Session × Type ($F_{1,69}$=6.35, p=.014); there were

no other main or interaction effects ($F \leq 3.85$, $p \geq .054$). For Session 1, the follow-up ANOVA for the Group × Type effect in the 3-way interaction showed a significant main effect of Type ($F_{1,138}=10.08$, $p=.002$), no significant main effect of Group ($F<0.01$, $p=.991$), and a significant 2-way interaction effect ($F_{1,138}=13.54$, $p<.001$), with independent groups t-tests showing HUD>CTL for FP ($t_{69}=-3.12$, $p=.003$) and CTL>HUD for TP ($t_{69}=2.43$, $p=.018$). There was a significant main effect of Type ($F_{1,138}=56.44$, $p<.001$) at Session 2, with a paired t-test showing TP>FP ($t_{70}=-6.17$, $p<.001$), and no other significant main or interaction effects ($F \leq 1.80$, $p \geq .180$). A repeated-measures ANOVA for the Session × Type effect in the 3-way interaction showed, for HUD, a significant main effect of Type ($F_{1,39}=5.08$, $p=.030$) and a significant 2-way interaction effect ($F_{1,39}=14.45$, $p<.001$), with paired t-tests showing Session 1>2 for FP ($t_{39}=3.00$, $p=.005$) and Session 2>1 for TP ($t_{39}=-3.31$, $p=.002$). The same test for CTL showed a significant effect of Type, TP>FP ($F_{1,30}=34.64$, $p<.001$). Paired t-tests comparing FP and TP frequencies showed, in HUD, no significant difference between pronoun types at Session 1 ($t_{39}=0.03$, $p=.972$), and TP>FP at Session 2 ($t_{39}=-3.82$, $p<.001$). The same tests in CTL showed TP>FP at both Session 1 ($t_{30}=-5.00$, $p<.001$) and Session 2 ($t_{30}=-5.47$, $p<.001$).

Additional term frequency analyses for control categories, which included place words and non-pronoun stop words, showed no significant effects (Figure S1).

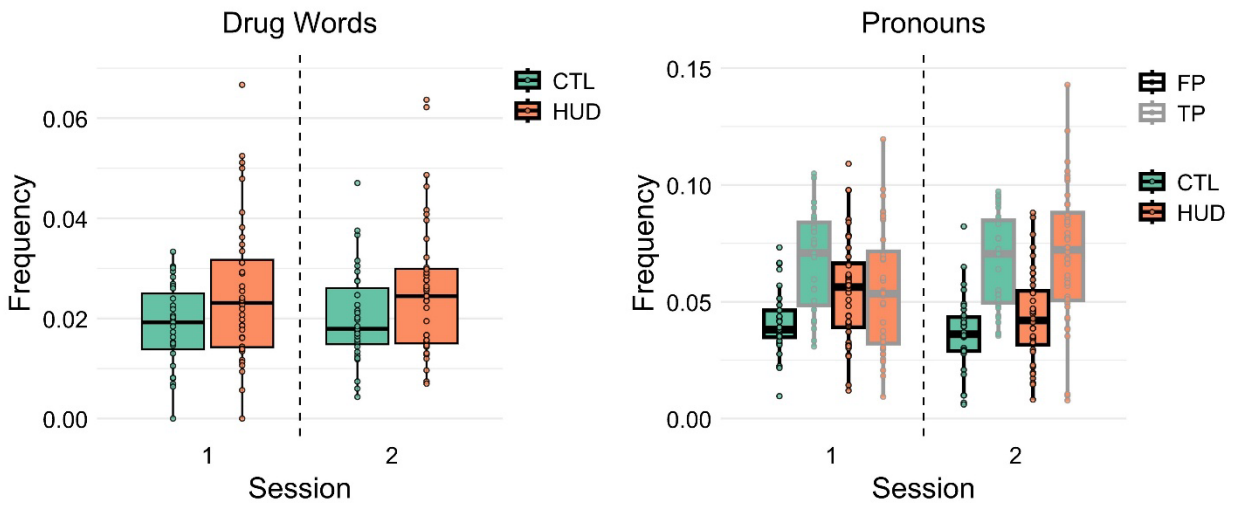

Figure 3. Term frequency analyses for selected categories of interest. Across both sessions, the HUD group referred to drugs significantly more than the CTL group, with no significant change between the sessions in either group. At Session 1, HUD participants used significantly more FP and less TP pronouns than CTL, with no significant differences between the groups at Session 2. Between the sessions, HUD participants showed decreased use of FP and increased use of TP pronouns, while CTL participants showed no significant changes in pronoun use. In the boxplots, the center line indicates the median, the box indicates the interquartile range, and whiskers extend to the most extreme points within 1.5x the interquartile range above and below the box.

Correlations with Outcome Measures

Language variables of interest (mean within- and between-group semantic similarity and drug word, FP pronoun, and TP pronoun frequencies) were inspected for correlations with craving measures that showed significant decreases at Session 2 (baseline, task-induced, and scene-induced craving) as well as depression and anxiety symptom scores in the HUD group. Spearman's correlation coefficients for these variables were assessed with their Session 1 measures and deltas (Session 2-Session 1). These tests yielded no significant results for the

baseline ($|r_s| \leq 0.28$, $p \geq .086$), task-induced ($|r_s| \leq 0.25$, $p \geq .124$), or scene-induced ($|r_s| \leq 0.14$, $p \geq .388$) craving measures, or for depression ($|r_s| \leq 0.25$, $p \geq .133$) and anxiety ($|r_s| \leq 0.29$, $p \geq .069$) symptoms.

Speech Representational Similarity to fMRI

To assess the extent to which variabilities in recall reflect neural activation patterns during movie viewing, we performed semantic-neural inter-subject representational similarity analyses (IS-RSA) on the recall LLM embeddings and fMRI data (Figure 4a). First, intersubject correlations (ISC) of the mean BOLD time series were inspected in eight cortical functional networks (control, default mode, dorsal attention, limbic, salience/ventral attention, somatomotor, temporoparietal, and visual), consolidated from the Schaefer atlas [24] to assess group similarities in neural activity during the movie viewing among HUD and CTL participants (Supplementary Materials and Methods). The dorsal attention, visual, and default mode networks showed lower group-level intersubject correlations in the HUD vs. CTL group at Session 1, with HUD ISCs increasing at Session 2 (Figure S5).

IS-RSAs were carried out in the same eight networks, with BOLD ISCs in dorsal attention and visual networks showing significantly greater representational similarity to movie recall embeddings in HUD compared to CTL at Session 1. No significant IS-RSA group differences were observed at Session 2, and the visual network representational similarity decreased between sessions in HUD (Table S5; see Table S6 for the baseline replication sample).

Similar comparisons were inspected in 416 individual regions of interest (ROIs), including 400 cortical parcellations of the same eight functional networks and 16 subcortical areas, to assess the specificity of the IS-RSA results. This analysis identified 14 ROIs in which the representational similarity was significantly higher in the HUD vs. CTL group ($p_{corr} < .05$) at

Session 1. These group differences were driven by significant positive correlations in HUD. Consistent with the network-level analyses, these significant ROIs encompassed cortical areas associated with the dorsal attention and visual networks, including the left dorsolateral prefrontal cortex/frontal eye field, bilateral superior parietal lobule, left postcentral gyrus, left medial and lateral occipital cortex and right occipital pole, and bilateral occipital fusiform gyrus. Significant CTL>HUD representational similarity was observed in one ROI in the left supramarginal gyrus. At Session 2, two ROIs in the right superior parietal lobule showed significant HUD>CTL representational similarity. The IS-RSA correlations in three ROIs located in the left dorsolateral prefrontal cortex/frontal eye field, left lateral occipital cortex, and left occipital fusiform gyrus decreased significantly in the HUD group from Session 1 to 2. Results from the ROI analyses are presented in Table 3 and Figure 4b (see Table S7 and Figure S6 for the baseline replication sample).

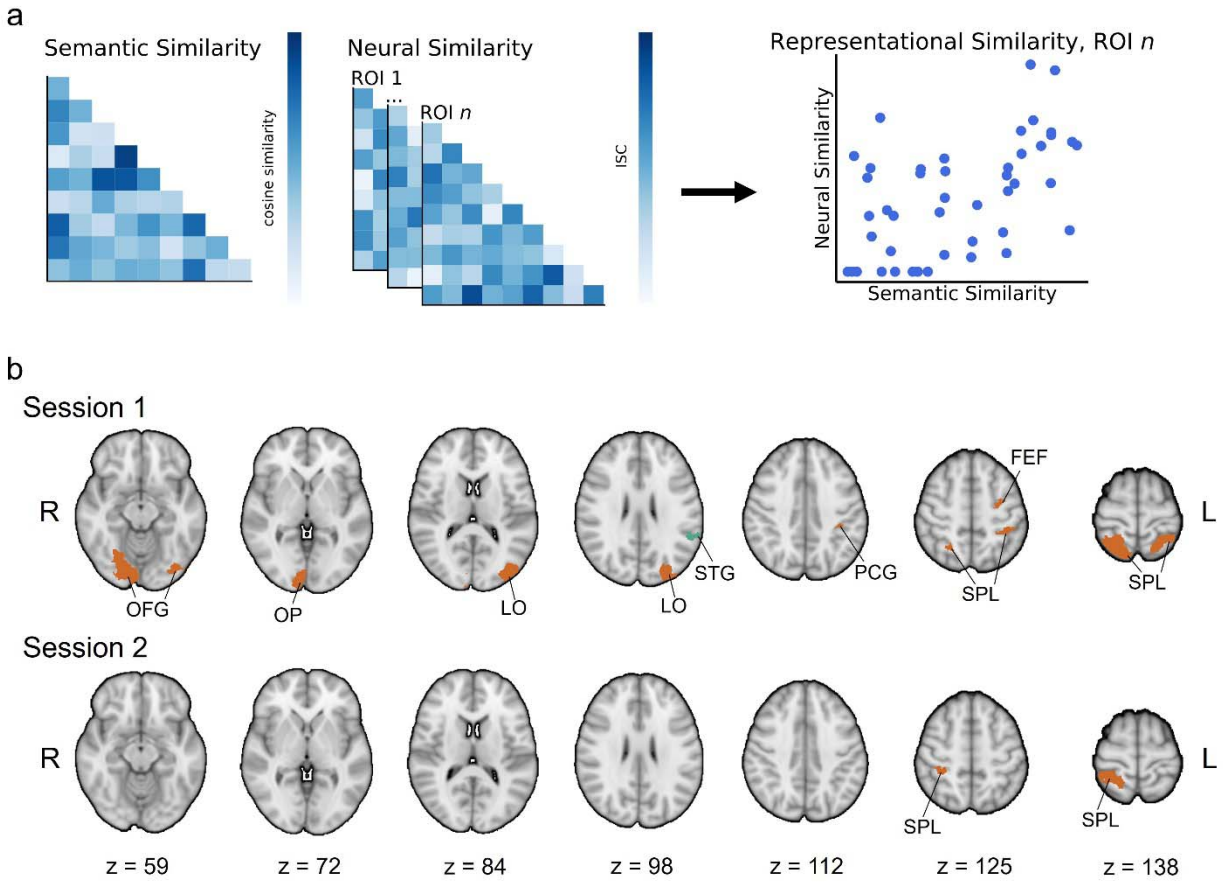

Figure 4. a: Overview of the RSA. Per session, correlations between one semantic similarity matrix (also see Figure 2a) and separate neural similarity matrices per region of interest (ROI) were computed using Mantel tests with Spearman's correlation coefficients. b: Anatomical distribution of the RSA results. Regions in which the Spearman's correlation was significantly higher in the HUD vs. CTL group (orange) or CTL vs. HUD (green), and which also showed a significant positive correlation in the HUD or CTL group (see Table 3), are shown here. OFG: occipital fusiform gyrus; OP: occipital pole; LO: lateral occipital cortex; STG: superior temporal gyrus; PCG: postcentral gyrus; SPL: superior parietal lobule; FEF: frontal eye field.

Table 3 – Statistics for semantic-neural representational similarity analysis

	Correlation Comparison, Z_{diff} (p_{corr})				Session 1, r_s (p)		Session 2, r_s (p	
Structure (Schaefer Atlas #)	Group: Session 1	Group: Session 2	Session: HUD	Session: CTL	CTL	HUD	CTL	HUI
HUD>CTL								
Left Occipital Fusiform Gyrus (3)[1]	**5.30 (<.001)**	-0.58 (>.999)	**5.83 (<.001)**	-	-0.09 (.755)	**0.24 (.008)**	-0.06 (.705)	-0.09 (.8
Left Lateral Occipital Cortex (10)[1]	**6.26 (<.001)**	**4.78 (.001)**	2.18 (>.999)	-	-0.16 (.868)	**0.24 (.015)**	-0.19 (.924)	0.11 (.1
Left Lateral Occipital Cortex (12)[1]	**5.89 (<.001)**	2.00 (>.999)	**5.29 (<.001)**	-	-0.05 (.615)	**0.32 (.004)**	-0.11 (.797)	0.02 (.4
Left Lateral Occipital Cortex (65)[1]	**5.73 (<.001)**	**3.87 (.046)**	3.37 (.308)	-	-0.06 (.650)	**0.30 (.004)**	-0.14 (.856)	0.11 (.1
Left Postcentral Gyrus (77)[1]	**5.26 (<.001)**	3.43 (.254)	1.42 (>.999)	-	-0.14 (.902)	**0.19 (.009)**	-0.11 (.864)	0.11 (.0
Left Superior Parietal Lobule (78)[1]	**5.52 (<.001)**	**3.94 (.033)**	1.05 (>.999)	-	-0.16 (.901)	**0.18 (.007)**	-0.13 (.853)	0.12 (.0
Left Superior Parietal Lobule (80)[1]	**5.97 (<.001)**	**4.34 (.006)**	1.83 (>.999)	-	-0.17 (.904)	**0.21 (.015)**	-0.17 (.924)	0.10 (.1
Left Frontal Eye Field (84)[1]	**3.89 (.041)**	0.05 (>.999)	**4.82 (<.001)**	-	-0.02 (.571)	**0.22 (.006)**	-0.06 (.695)	-0.05 (.8
Right Occipital Fusiform Gyrus (203)[1]	**4.31 (.007)**	2.68 (>.999)	3.48 (.206)	-	-0.07 (.682)	**0.20 (.043)**	-0.17 (.893)	>-0.01 (.
Right Occipital Fusiform Gyrus (204)[1]	**4.34 (.006)**	2.45 (>.999)	3.50 (.190)	-	-0.04 (.603)	**0.23 (.006)**	-0.12 (.836)	0.03 (.6
Right Occipital Pole (207)[1]	**4.62 (.002)**	1.97 (>.999)	3.65 (.107)	-	-0.07 (.694)	**0.22 (.002)**	-0.12 (.831)	>-0.01 (.
Right Lateral Occipital Cortex (210)	**4.24 (.009)**	2.56 (>.999)	1.28 (>.999)	-	-0.08 (.720)	0.18 (.054)	-0.05 (.640)	0.11 (.1
Right Occipital Pole (211)	**5.06 (<.001)**	3.28 (.430)	0.12 (>.999)	-	-0.18 (.924)	0.14 (.074)	-0.07 (.725)	0.13 (.0
Right Lateral Occipital Cortex (212)	**4.41 (.004)**	**4.31 (.006)**	0.54 (>.999)	-	-0.12 (.799)	0.15 (.078)	-0.15 (.859)	0.12 (.1
Right Superior Parietal Lobule (271)[1]	**5.22 (<.001)**	2.29 (>.999)	3.29 (.421)	-	-0.05 (.638)	**0.27 (.003)**	-0.06 (.650)	0.02 (.2
Right Superior Parietal Lobule (272)[1]	**3.87 (.045)**	2.50 (>.999)	2.97 (>.999)	-	-0.05 (.656)	**0.19 (.025)**	-0.14 (.870)	0.02 (.4
Right Superior Parietal Lobule (278)[2]	2.62 (>.999)	**5.24 (<.001)**	-1.27 (>.999)	-	-0.04 (.601)	0.12 (.064)	-0.13 (.870)	**0.19 (.0**
Right Superior Parietal Lobule (280)[1,2]	**5.27 (<.001)**	**5.61 (<.001)**	0.04 (>.999)	-	-0.12 (.813)	**0.21 (.015)**	-0.14 (.879)	**0.21 (.0**
CTL>HUD								
Left Supramarginal Gyrus (87)[1]	**-4.49 (.003)**	-1.34 (>.999)	-	2.75 (>.999)	**0.29 (.003)**	0.01 (.387)	0.11 (.119)	0.03 (.3

Notes: Table presents statistics for structures that met criteria for significance in representational similarity analysis of recall semantic similarity and BOI ISC during movie viewing in the replication sample. Comparisons of the correlation coefficients between groups/within each session, and between sessio within the groups, were performed using Fisher Z testing (Z_{diff}) with Bonferroni-corrected p values, shown in the Correlation Comparison colun Spearman r (r_s) and uncorrected p-values derived from Mantel tests are shown for each group. Bold values indicate statistically significant tests (p<.0 Regions that showed a significant group difference in the correlation coefficient and a significant RSA correlation within a group are denoted [1] for Sessi 1 and [2] for Session 2.

Discussion

This study demonstrates, for the first time, how naturalistic language could provide a sensitive measure of altered neurocognitive function in HUD. Using contextual analysis of free recall following a drug-related movie, we identified distinct linguistic signatures that differentiated HUD participants from healthy controls and tracked changes during treatment. At baseline, HUD participants exhibited greater intersubject variability, characterized by increased drug-related and self-referential, and reduced socially directed language; these features partially normalized with three months of inpatient treatment. Importantly, semantic variability aligned with individual differences in brain responses during movie viewing, particularly within dorsal attention and visual networks, suggesting that attentional encoding dynamics shape later verbal recall. Semantic-neural correspondences were stable across time points in superior parietal regions of the dorsal attention network. Critically, the observed semantic and semantic-neural effects at baseline were replicated in an independent group of HUD participants, underscoring the robustness of spontaneous recall as a marker of underlying neurocognitive functioning in addiction.

Prior work has shown that quantifiable features of language can be leveraged for diagnostic classification, subgrouping, and outcome prediction in psychiatric disorders, revealing distinct linguistic signatures associated with disease states and treatment outcomes [3]. Notably, most prior studies in addiction have relied on natural language from secondary sources such as clinician notes [25] or user-generated content on social media, including anonymous posts in online support groups [4–7]. Here, we used an experimental approach in well-phenotyped inpatients with HUD to elicit natural language tied to a salient, naturalistic context. This design revealed semantic features that not only distinguished HUD from controls but also covaried with the

participants' neural activity. By demonstrating that linguistic variability corresponds to functional variability in large-scale attentional systems, our results extend prior work by showing that naturalistic speech captures meaningful individual differences in brain processing, even when separated in time from the event being recalled. These findings establish natural language as a mechanistic marker of addiction and neurocognitive adaptation during recovery, complementing traditional imaging approaches.

The specific linguistic markers observed, including drug-focused language and shifts in pronoun use, map onto core cognitive processes disrupted in addiction, including biased attention, self-referential processing, and social cognition. For example, greater use of drug words suggests biased attentional allocation towards drug reinforcers [26], consistent with models of disrupted salience attribution in substance use disorders. Self-referential language may reflect heightened interpersonal distress [27], or here personal identification with the drug-themed content, whereas increased use of third-person pronouns could signal more externally directed or socially oriented processing. Similar associations have been observed in other psychiatric conditions [28–30], suggesting that these features could capture transdiagnostic dimensions of cognitive-affective function. Future work should test the generalizability of these markers across contexts, including other drug-themed movies or drug-neutral stimuli, different recall tasks, and naturalistic speech obtained in everyday settings. Such studies will clarify the extent to which these features can predict long-term clinical outcomes and relapse risk.

Semantic-neural representational similarity in this movie recall task emerged in dorsal attention and visual network regions central to the allocation of motivated attention, including superior parietal lobule, frontal eye field, and occipital cortex. Attentional bias towards drug cues is fundamental to addiction, representing a critical target for relapse prevention and recovery

[31,32]. While the visual system's role in drug cue processing has received little attention, most cue reactivity studies report drug-biased activations in visual areas [33], potentially related to feedback mechanisms through which attention and reward-related processes in higher cortical regions, informed by prior experiences, signal to adjust response preferences in early sensory areas towards salient or personally relevant stimuli [33–35]. Although cue-reactivity studies typically emphasize valuation and reward circuitry, our results highlight the contribution of attentional and sensory systems to the encoding of complex, salient contexts. The absence of semantic-neural correspondence in orbitofrontal cortex, despite its role in our prior movie-viewing work [23], suggests that delayed recall tasks may preferentially capture longer timescale attentional dynamics (e.g., attentional guidance mediated by prior relevant experiences [36–38]) rather than immediate value-based encoding. Importantly, the reliability of the observed effect across both sessions, and a separate group of subjects, indicates that spontaneous recall provides a robust index of brain encoding processes, positioning language as a practical and reproducible marker of neurocognitive function in addiction.

Despite encouraging findings, this study has limitations. We did not identify associations between linguistic features and clinical measures such as craving, depression, or anxiety, potentially due to the relatively short (~3 month) follow-up interval; it is also possible that the former capture more implicit processes than the latter, all based on self-evaluative declarative processes. Longer-term longitudinal designs, with denser sampling across treatment and post-discharge, will be needed to establish the predictive validity of linguistic markers. In addition, the present findings were obtained in a specific drug-related narrative context; future studies should test whether speech markers generalize to neutral or other personally relevant contexts and whether they can track drug use and/or relapse risk in real-world environments.

To conclude, these results establish unstructured verbal recall, a scalable and ecologically valid tool, for probing the neural encoding of salient contexts in addiction. By linking spontaneous speech to distributed attentional systems, we show that natural language provides a robust behavioral marker of specific neurocognitive functions and recovery. This approach opens new opportunities for mechanistic studies of addiction and related psychiatric conditions, with potential applications for digital phenotyping, treatment monitoring, and outcome prediction.

Materials and Methods

Participants

Forty individuals with HUD and 31 CTL participants were recruited through a clinical trial evaluating intervention-related recovery outcomes in opiate addiction (NCT04112186), for which therapy-specific results will be reported separately. Participants completed all movie task procedures twice, during an initial baseline visit (Session 1) and approximately 15 weeks later (Session 2). During the inter-scan interval, HUD participants completed an 8-week group therapy program in addition to their regular inpatient treatment (see below). For the analyses that involved fMRI data, one CTL participant was excluded due to a scanner malfunction that resulted in data loss but did not disrupt the movie viewing, and five HUD participants were excluded due to excessive head motion (i.e., 30 CTL/35 HUD participants were included in these analyses). An additional 32 HUD participants and 2 CTL participants who only completed Session 1 were retained as part of a replication dataset for the baseline analyses described below (see Supplementary Methods and Results for details).

All HUD participants were recruited as inpatients enrolled in a treatment program that included prescription opioid medications for opioid use disorder (methadone: n=37, buprenorphine/naloxone: n=3) as well as regular therapy counseling and relapse prevention, trauma, and anger management courses, at a collaborating substance use rehabilitation facility in the New York City metropolitan area (Samaritan Daytop Village, Queens, NY). Age- and gender-matched CTL participants were recruited through advertisements and word-of-mouth in the surrounding community.

Eligibility and participant characteristics were assessed during an initial screening with a comprehensive clinical interview administered by trained research coordinators supervised by

clinical psychologists. Diagnostic evaluations included the Mini International Neuropsychiatric Interview, 7th Edition [39] and the Addition Severity Index, 5th Edition [40]. All HUD participants met criteria for opiate use disorder with heroin as their primary substance of choice, and CTL participants did not meet criteria for any current or past substance use disorder. Craving and withdrawal symptoms in HUD participants were assessed with the Heroin Craving Questionnaire (HCQ), a modified version of the Cocaine Craving Questionnaire [41], and the Short Opiate Withdrawal Scale (SOWS) [42], and their drug dependence severity was assessed with the Severity of Dependence Scale (SDS) [43]. Methadone or buprenorphine use was confirmed through urine toxicology at all screening and study visits. For all participants, current or past nicotine use and alcohol dependence were assessed with the Fagerström Test for Nicotine Dependence (FTND) [44] and the Short Michigan Alcohol Screening Test (SMAST) [45]. A brief medical examination including heart rate, blood pressure, alcohol concentration (with saliva test strips), breath carbon monoxide, and pregnancy testing when applicable (with urine test strips) was administered by trained research coordinators during each visit. Estimated verbal and nonverbal IQ were assessed with the reading subtest of the Wide Range Achievement Test-3 [46] and the Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence [47], respectively. Other neuropsychological measures, which were acquired during study visits conducted before participants began the experimental protocol, included assessments for symptoms of depression (Beck's Depression Inventory, BDI) [48] in the past two weeks and anxiety (Beck's Anxiety Inventory, BAI) [49] in the past month. Group comparisons on demographic, neuropsychological, and substance use characteristics were performed using two-sample Wilcoxon signed-rank tests for continuous variables and chi-squared tests for categorical variables in R version 6.3.1.

Comorbidities in the HUD participants, which were in partial or sustained remission, included major depression (n=15), anxiety disorders (n=7), post-traumatic stress disorder (n=10), and other substance use disorders (alcohol: n=10; cannabis: n=4; cocaine: n=13; other stimulants: n=2; tranquilizers: n=6). Three CTL participants met criteria for past depression, and one for past anxiety. All participants were otherwise medically healthy.

Experimental Protocol

During both experiment sessions, participants viewed the first 17 minutes and 3 seconds of the movie *Trainspotting* with fMRI acquisition as described in our previous publication [23]. Scans were acquired on a Siemans 3T Skyra (Siemans, Erlangen, Germany) with a 32-channel head coil as part of a 2-hour MRI protocol that included other fMRI tasks [20,50,51] and T1-, T2-, and diffusion-weighted scanning [52–55]. Details on the scanning procedures and MRI preprocessing steps can be found in our previous publication [23] and in the Supplementary Methods and Results.

Within 45 minutes of completing the MRI procedures, participants were instructed to recall the movie narrative and describe their experience of watching it while their speech was recorded. A research coordinator gave instructions as follows: “Please verbally recall the movie you just watched, in order, with as much detail as possible. The details are more important than the order. While recounting the movie, please also describe any thoughts, feelings, or emotions you may have experienced. Please tell us what these thoughts were and indicate the specific items in each of the scenes that made you have these feelings, and how strong your emotions were.” Participants performed the recall task uninterrupted, and no minimum or maximum time limit was imposed. Afterwards, participants completed a quality assurance questionnaire

(Supplementary Methods and Results) to assess their memory and comprehension of the movie, indicate whether they had seen it before, and rate their attention and understanding of the movie and the audio and visual quality; these measures did not differ significantly between the groups (Table S1; see Table S2 for the baseline replication sample).

To investigate potential relationships between the dependent variables and craving, a hallmark symptom of addiction that is a strong predictor of future relapse [56–58], participants were asked to rate their current heroin craving immediately before (baseline craving) and after (task-induced craving) the movie viewing ("Please rate how strong your desire for heroin is on a scale of 0-9"). Following the speech task, participants also rated their feelings of craving (scale of 0-9) related to specific scenes (scene-induced craving), sampled from 35 three-second video clips taken at regular 30-second intervals from the entire movie segment. Session comparisons for these craving and other repeated outcome measures of interest in HUD participants were performed using paired Wilcoxon signed-rank tests (Table 2).

Speech Preprocessing

Audio processing and text analysis were performed in Python version 3.10.11. Recordings were transcribed to text using OpenAI's Whisper automated speech recognition model [59]. Transcribed texts underwent manual quality assurance processing for accuracy, including removal of interviewer speech and irrelevant participant speech (e.g., asking for clarification on the task instructions). The transcripts were then embedded into 768-dimensional vectors using the all-distilroberta-v1 pretrained model from the SentenceTransformers framework for Bidirectional Encoder Representations from Transformers (BERT) sentence encoding [60]. Briefly, sentence transformers are a deep learning model pretrained on a large

corpus of sentence pairs with varying similarities in meaning to capture latent contextual features in short strings of text, which are represented by an output vector of "embeddings." Comparisons between pairs of embeddings vectors can be obtained using metrics such as the cosine similarity, which measures the closeness (or semantic similarity) between two vectors in embeddings space (in this case as the cosine of the angle between the two vectors, such that values closer to -1 indicate greater semantic distance and values closer to 1 indicate greater semantic similarity). Since most of the transcripts exceeded the 512-token limit (mean/standard deviation words per sample: 650/429.8; range: 83-2,479, with approximately one token per 0.75 words), samples were divided into windows of 40 words, equal to approximately half the length of the shortest transcript; an embeddings vector was extracted for each window, and these were averaged across all the windows from each transcript, resulting in one 768-dimensional vector per sample. Transcripts that were not evenly divisible by the window size were truncated by the length of the remainder.

Classification Analysis

To test the decoding of group and session from the 142 transcripts (71 participants with two sessions each), dimensionality reduction was performed on the embeddings vectors with principal component analysis using scikit-learn [61]. The first 25 principal components, representing a cumulative variance ratio of 75.4%, were selected for classification to avoid overfitting the classifier while retaining a substantial proportion of the overall variability in the data. These steps are in line with previous work that found principal component analysis to perform better than other dimensionality reduction techniques in handling contextual embeddings for mental health classification tasks when the number of embeddings dimensions

exceeds the sample size [62]. Classifier performance on the principal components, with categorical labels for group and session, was assessed with the scikit-learn function permutation_test_score using a logistic regression classifier with L2 regularization, $\lambda=1$, and 10-fold cross validation, with 1000 permutations of shuffled labels used to generate a null accuracy distribution. True accuracy scores above the 95th percentile ($p<.05$) of the null distribution were considered statistically significant.

Semantic Similarity Analysis

To assess the semantic similarity between participants' movie recall, cosine similarity values were computed between every pair of embeddings vectors within the same session using the scikit-learn cosine similarity function (Figure 2a). For statistical comparisons, the pairwise cosine similarity between each sample and each other sample from the same and other group were averaged separately to compute a mean within-group (CTL-CTL or HUD-HUD) and between-group (HUD-CTL) semantic similarity value for each participant in both sessions. Additionally, a within-subject/between-sessions similarity value was computed for each participant using the cosine similarity of their embeddings vectors from Sessions 1 and 2. The averaged cosine similarity values were evaluated statistically using a 3 (Comparison: CTL-CTL, HUD-HUD, HUD-CTL) × 2 (Session: 1, 2) mixed ANOVA, and an independent groups t-test was used to compare the within-subject similarity between the groups. Statistical tests were performed in R version 4.4.2 (R Core Team).

Term Frequency Analysis

To quantify term frequencies for categories of interest, text samples underwent lemmatization and expansion of contractions using Natural Language Toolkit (NLTK) modules in Python [64], and word counts were extracted for the whole sample and specific term categories as described in the Supplementary Materials and Methods. Frequencies for each category were calculated by dividing the sum of the individual word counts within a category by the total word count for the sample. Table S3 presents categories and word lists used for this analysis, which included drug-related words and first-person (FP) and third-person (TP) pronouns as well as control categories. Group and session comparisons on term frequencies were analyzed using a 2 (Group: HUD, CTL) × 2 (Session: 1, 2) × 2 (Type: FP, TP) mixed ANOVA for pronouns and 2 (Group) × 2 (Session) mixed ANOVAs for other categories. Statistical tests were performed in R version 4.4.2 (R Core Team).

Semantic-Neural Inter-subject Representational Similarity Analysis

To assess the representation of BOLD signal patterns during movie viewing in semantic patterns from the verbal recall of the movie, we performed inter-subject representational similarity analysis (IS-RSA) on pairwise semantic and neural similarity matrices [65–67]. Semantic similarity matrices consisted of the pairwise cosine similarity values computed between participants' embeddings vectors (Figure 2A). Neural similarity matrices consisted of Pearson correlation coefficients (intersubject correlations) computed between participants' z-scored BOLD time series data. These were extracted separately for 400 cortical regions of interest (ROIs) from the Schaefer 17-network functional atlas [24] and 16 subcortical ROIs from the Harvard-Oxford subcortical structural atlas (RRID:SCR_001476, provided by FSL v.6.0.3, FMRIB, Oxford, UK), resulting in 416 separate neural similarity matrices in each session.

Semantic-neural IS-RSA was performed using the lower triangle of the semantic similarity matrix with each of the lower triangles of the 416 neural similarity matrices, separately, with Mantel testing [68] using the "mantel" package for Python (https://github.com/jwcarr/mantel). Specifically, we computed Spearman's rank correlation coefficients along with 10,000 permutations of shuffled labels to generate a null distribution for hypothesis testing. Since only significant positive correlations, indicating co-variability between participants in their verbal recall and BOLD responses, could be meaningfully interpreted in this analysis, hypothesis tests were performed one-tailed on the upper tail.

Comparisons of the Spearman's correlation coefficients (r_s) between groups and sessions for each ROI were performed using a Fisher r -to-z transformation: r_s values were z-transformed and the absolute difference (Z_{diff}) computed as the difference between the z-scores divided by the standard error of the difference. Test statistics were assigned p-values using the Normal Cumulative Distribution Function from SciPy [69], scipy.stats.norm.cdf, followed by Bonferroni correction (p_{corr}) for 416 separate tests. Regions showing significant ($p_{corr}<.05$) group differences between the correlation coefficients in either session were further examined for significant session differences and r_s within the groups.

References

1. Malgaroli, M., Hull, T. D., Zech, J. M. & Althoff, T. Natural language processing for mental health interventions: a systematic review and research framework. *Transl Psychiatry* **13**, 309 (2023).
2. Zhang, T., Schoene, A. M., Ji, S. & Ananiadou, S. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine* vol. 5 Preprint at https://doi.org/10.1038/s41746-022-00589-7 (2022).
3. Malgaroli, M., Hull, T. D., Zech, J. M. & Althoff, T. Natural language processing for mental health interventions: a systematic review and research framework. *Transl Psychiatry* **13**, 309 (2023).
4. Liu, T. *et al*. Linguistic predictors from Facebook postings of substance use disorder treatment retention versus discontinuation. *Am J Drug Alcohol Abuse* **48**, 573–585 (2022).
5. Yang, G., King, S. G., Lin, H. M. & Goldstein, R. Z. Emotional Expression on Social Media Support Forums for Substance Cessation: Observational Study of Text-Based Reddit Posts. *J Med Internet Res* **25**, (2023).
6. Curtis, B. *et al.* AI-based analysis of social media language predicts addiction treatment dropout at 90 days. *Neuropsychopharmacology* **48**, (2023).
7. Kornfield, R., Toma, C. L., Shah, D. V., Moon, T. J. & Gustafson, D. H. What Do You Say Before You Relapse? How Language Use in a Peer-to-peer Online Discussion Forum Predicts Risky Drinking among Those in Recovery. *Health Commun* **33**, 1184–1193 (2018).
8. Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D. & Hasson, U. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences* **111**, (2014).
9. Hasson, U., Egidi, G., Marelli, M. & Willems, R. M. Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition* **180**, 135–157 (2018).
10. ELLIS, N. C. Essentials of a Theory of Language Cognition. *The Modern Language Journal* **103**, 39–60 (2019).
11. Chan, C. C. *et al*. Emergence of Language Related to Self-experience and Agency in Autobiographical Narratives of Individuals With Schizophrenia. *Schizophr Bull* **49**, (2023).
12. Cox, D. J., Garcia-Romeu, A. & Johnson, M. W. Predicting changes in substance use following psychedelic experiences: natural language processing of psychedelic session narratives. *American Journal of Drug and Alcohol Abuse* **47**, (2021).
13. Agurto, C. *et al.* Speak and you shall predict: evidence that speech at initial cocaine abstinence is a biomarker of long-term drug use behavior. *Biol Psychiatry* (2025) doi:10.1016/j.biopsych.2025.01.009.
14. de Wit, H., Epstein, D. H. & Preston, K. L. Does human language limit translatability of clinical and preclinical addiction research? *Neuropsychopharmacology* **43**, 1985–1988 (2018).
15. Garnett, M., Miniño, A., Joyce, M., Driscoll, A. & Valenzuela, C. *Drug Overdose Deaths in the United States, 2003–2023*. (2024) doi:10.15620/cdc/170565.

16. Jones, C. M., Han, B., Baldwin, G. T., Einstein, E. B. & Compton, W. M. Use of Medication for Opioid Use Disorder Among Adults With Past-Year Opioid Use Disorder in the US, 2021. *JAMA Netw Open* **6**, (2023).
17. Williams, A. R., Samples, H., Crystal, S. & Olfson, M. Acute Care, Prescription Opioid Use, and Overdose Following Discontinuation of Long-Term Buprenorphine Treatment for Opioid Use Disorder. *American Journal of Psychiatry* **177**, (2020).
18. Goldstein, R. Z. & Volkow, N. D. Dysfunction of the prefrontal cortex in addiction: Neuroimaging findings and clinical implications. *Nature Reviews Neuroscience* vol. 12 Preprint at https://doi.org/10.1038/nrn3119 (2011).
19. Edwards, S. & Koob, G. F. Escalation of drug self-administration as a hallmark of persistent addiction liability. *Behavioural Pharmacology* vol. 24 Preprint at https://doi.org/10.1097/FBP.0b013e3283644d15 (2013).
20. Huang, Y. *et al.* Association of Cortico-Striatal Engagement During Cue Reactivity, Reappraisal, and Savoring of Drug and Non-Drug Stimuli With Craving in Heroin Addiction. *Am J Psychiatry* **181**, (2024).
21. Zilverstand, A., Huang, A. S., Alia-Klein, N. & Goldstein, R. Z. Neuroimaging Impaired Response Inhibition and Salience Attribution in Human Drug Addiction: A Systematic Review. *Neuron* vol. 98 Preprint at https://doi.org/10.1016/j.neuron.2018.03.048 (2018).
22. Bossert, J. M., Marchant, N. J., Calu, D. J. & Shaham, Y. The reinstatement model of drug relapse: recent neurobiological findings, emerging research topics, and translational research. *Psychopharmacology (Berl)* **229**, 453–476 (2013).
23. Kronberg, G. *et al.* Shared orbitofrontal dynamics to a drug-themed movie track craving and recovery in heroin addiction. *Brain* **148**, 1778–1788 (2025).
24. Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol* **106**, 1125–1165 (2011).
25. Tassinari, D. L. *et al.* Artificial Intelligence-driven and technological innovations in the diagnosis and management of substance use disorders. *International Review of Psychiatry* **37**, 52–58 (2025).
26. Goldstein, R. Z., Woicik, P. A., Lukasik, T., Maloney, T. & Volkow, N. D. Drug fluency: A potential marker for cocaine use disorders. *Drug Alcohol Depend* **89**, 97–101 (2007).
27. Zimmermann, J., Wolf, M., Bock, A., Peham, D. & Benecke, C. The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems. *J Res Pers* **47**, 218–225 (2013).
28. Elleuch, D., Chen, Y., Luo, Q. & Palaniyappan, L. Speaking of yourself: A meta-analysis of 80 years of research on pronoun use in schizophrenia. *Schizophr Res* **279**, 22–30 (2025).
29. Huang, R. *et al.* Exploring the Role of First-Person Singular Pronouns in Detecting Suicidal Ideation: A Machine Learning Analysis of Clinical Transcripts. *Behavioral Sciences* **14**, 225 (2024).
30. Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H. & Wolf, M. First□person Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients. *Clin Psychol Psychother* **24**, 384–391 (2017).
31. Escudero, B., Arias Horcajadas, F. & Orio, L. Changes of attentional bias in patients with alcohol use disorder during abstinence: A longitudinal study. *Addictive Behaviors* **157**, 108098 (2024).

32. Parvaz, M. A. *et al.* Attention bias modification in drug addiction: Enhancing control of subsequent habits. *Proceedings of the National Academy of Sciences* **118**, (2021).
33. Hanlon, C. A., Dowdle, L. T., Naselaris, T., Canterberry, M. & Cortese, B. M. Visual cortex activation to drug cues: A meta-analysis of functional neuroimaging papers in addiction and substance abuse literature. *Drug Alcohol Depend* **143**, 206–212 (2014).
34. Vossel, S., Geng, J. J. & Fink, G. R. Dorsal and Ventral Attention Systems. *The Neuroscientist* **20**, 150–159 (2014).
35. Krauzlis, R. J., Lovejoy, L. P. & Zénon, A. Superior Colliculus and Visual Spatial Attention. *Annu Rev Neurosci* **36**, 165–182 (2013).
36. Xu, Y. Parietal-driven visual working memory representation in occipito-temporal cortex. *Current Biology* **33**, 4516-4523.e5 (2023).
37. Patai, E. Z., Doallo, S. & Nobre, A. C. Long-term Memories Bias Sensitivity and Target Selection in Complex Scenes. *J Cogn Neurosci* **24**, 2281–2291 (2012).
38. Nobre, A. C. & Stokes, M. G. Premembering Experience: A Hierarchy of Time-Scales for Proactive Attention. *Neuron* **104**, 132–146 (2019).
39. Sheehan, D. V. *et al.* The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. in *Journal of Clinical Psychiatry* vol. 59 (1998).
40. McLellan, A. T. *et al.* The fifth edition of the addiction severity index. *J Subst Abuse Treat* **9**, 199–213 (1992).
41. Tiffany, S. T., Singleton, E., Haertzen, C. A. & Henningfield, J. E. The development of a cocaine craving questionnaire. *Drug Alcohol Depend* **34**, 19–28 (1993).
42. Gossop, M. The development of a short opiate withdrawal scale (SOWS). *Addictive Behaviors* **15**, (1990).
43. Gossop, M. *et al.* The Severity of Dependence Scale (SDS): psychometric properties of the SDS in English and Australian samples of heroin, cocaine and amphetamine users. *Addiction* **90**, 607–614 (1995).
44. Heatherton, T. F., Kozlowski, LynnT., Frecker, R. C. & Fagerstrom, K. □O. The Fagerström Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Br J Addict* **86**, 1119–1127 (1991).
45. Selzer, M. L., Vinokur, A. & Van Rooijen, L. A self administered short Michigan Alcoholism Screening Test (SMAST). *J.STUD.ALCOHOL* **36**, (1975).
46. Jastak, S. & Wilkinson, G. Wide Range Achievement Test—Revised 3. *Wilmington, DE: Jastak Associates* (1993).
47. Wechsler, D. *Manual for the Wechsler Abbreviated Intelligence Scale (WASI). WASI* (1999).
48. Beck, A. T., Steer, R. A. & Brown, G. K. Manual for the Beck depression inventory-II. *San Antonio, TX: Psychological Corporation* (1996).
49. Beck, A. T. & Steer, R. A. Manual for the Beck Anxiety Inventory. *Behaviour research and therapy* vol. 37 Preprint at (1990).
50. Ceceli, A. O., King, S. G., McClain, N., Alia-Klein, N. & Goldstein, R. Z. The Neural Signature of Impaired Inhibitory Control in Individuals with Heroin Use Disorder. *Journal of Neuroscience* **43**, (2022).
51. Ceceli, A. O. *et al.* Recovery of anterior prefrontal cortex inhibitory control after 15 weeks of inpatient treatment in heroin use disorder. *Nature Mental Health* **2**, 694–702 (2024).

52. King, S. G. *et al.* Prefrontal-habenular microstructural impairments in human cocaine and heroin addiction. *Neuron* **110**, (2022).
53. Gaudreault, P. O., King, S. G., Malaker, P., Alia-Klein, N. & Goldstein, R. Z. Whole-brain white matter abnormalities in human cocaine and heroin use disorders: association with craving, recency, and cumulative use. *Mol Psychiatry* **28**, (2023).
54. Gaudreault, P.-O. *et al.* Frontal White Matter Changes and Craving Recovery in Inpatients With Heroin Use Disorder. *JAMA Netw Open* **7**, e2451678 (2024).
55. Ceceli, A. O. *et al.* Common and distinct fronto-striatal volumetric changes in heroin and cocaine use disorders. *Brain* **146**, 1662–1671 (2023).
56. Ellis, J. D. *et al.* Intra-individual variability and stability of affect and craving among individuals receiving medication treatment for opioid use disorder. *Neuropsychopharmacology* **47**, 1836–1843 (2022).
57. Burgess-Hull, A. J., Panlilio, L. V., Preston, K. L. & Epstein, D. H. Trajectories of craving during medication-assisted treatment for opioid-use disorder: Subtyping for early identification of higher risk. *Drug Alcohol Depend* **233**, 109362 (2022).
58. Vafaie, N. & Kober, H. Association of Drug Cues and Craving with Drug Use and Relapse: A Systematic Review and Meta-analysis. *JAMA Psychiatry* vol. 79 Preprint at https://doi.org/10.1001/jamapsychiatry.2022.1240 (2022).
59. Radford, A. *et al.* Robust Speech Recognition via Large-Scale Weak Supervision. in *Proceedings of Machine Learning Research* vol. 202 (2023).
60. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (2019). doi:10.18653/v1/d19-1410.
61. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, (2011).
62. Ganesan, A. V *et al.* Empirical Evaluation of Pre-trained Transformers for Human-Level NLP: The Role of Sample Size and Dimensionality. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4515–4532 (Association for Computational Linguistics, 2021).
63. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.
64. Wagner, W. *Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit*. *Language Resources and Evaluation* vol. 44 (2010).
65. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front Syst Neurosci* (2008) doi:10.3389/neuro.06.004.2008.
66. Popal, H., Wang, Y. & Olson, I. R. A Guide to Representational Similarity Analysis for Social Neuroscience. *Soc Cogn Affect Neurosci* **14**, 1243–1253 (2019).
67. Dimsdale-Zucker, H. R. & Ranganath, C. Representational Similarity Analyses. in 509–525 (2018). doi:10.1016/B978-0-12-812028-6.00027-6.
68. Mantel, N. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Res* **27**, (1967).
69. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).