Tracking the behavioral and neural dynamics of semantic representations through negation

Arianna Zuanazzi[1*], Pablo Ripollés[1,2,3], Wy Ming Lin[4], Laura Gwilliams[5], Jean-Rémi King[1,6†], David Poeppel[1,3,7†]

1 Department of Psychology, New York University, New York, NY, USA.
2 Music and Audio Research Lab (MARL), New York University, New York, NY, USA.
3 Center for Language, Music and Emotion (CLaME), New York University, New York, NY, USA.
4 Hector Research Institute for Education Sciences and Psychology, University of Tübingen, Tübingen, Germany.
5 Department of Neurological Surgery, University of California, San Francisco, CA, USA.
6 Ecole Normale Supérieure, PSL University, Paris, France.
7 Ernst Strüngmann Institute for Neuroscience, Frankfurt, Germany.

†These authors contributed equally to this work.
*Corresponding author. Email: az1864@nyu.edu.

Abstract

Combinatoric linguistic operations underpin human language processes, but how meaning changes over time is not well understood. We address this puzzle by exploiting the ubiquitous function of negation. We track the online effects of negation ("not") and intensifiers ("really") on the representation of scalar adjectives (e.g., "good") in parametrically designed behavioral and neurophysiological (MEG) experiments. The behavioral data show that participants first interpret negated adjectives as affirmative and then modify their interpretation towards, but never exactly as, the opposite meaning. Decoding analyses of neural activity further reveal that negation does not *invert* the representation of adjectives (i.e., "not bad" represented as "good") but rather *mitigates* their representation, at early lexical-semantic processing stages. This putative suppression mechanism of negation is supported by increased synchronization of beta-band neural activity in sensorimotor areas. The analysis of negation provides a steppingstone to understand how the human brain represents changes of meaning over time.

Introduction

A hallmark of language processing is that we combine elements of the stored inventory - informally speaking, words - and thereby flexibly generate new meanings or change current meanings. The final representations derive in systematic ways from the combination of individual pieces. The composed meanings can relate in relatively straightforward ways to the building blocks (e.g., "the cat sat on the mat") or stem from more subtle inferential processes (e.g., "this theory is not even wrong"). A mechanistic understanding of the underlying processes requires characterization of how meaning representations are constructed in real time. There has been steady progress and productive debate on syntactic structure building [1–6]. In contrast, how novel semantic configurations are represented over time is less widely investigated. In the experimental approach pursued here, we build on the existing literature on precisely controlled *minimal* linguistic environments [7,8]. We deploy a new, simple parametric experimental paradigm that capitalizes on the powerful role that *negation* plays in shaping semantic representations of words. While negation is undoubtfully a complex linguistic operation that can affect comprehension as a function of other linguistic factors (such as discourse and pragmatics [9–11]), our investigation specifically focuses on how negation operates in phrasal structures. Combining behavioral and neurophysiological data, we show how word meaning is (and is not) modulated in controlled contexts that contrast affirmative (e.g., "really good") and negated (e.g., "not good") phrases. The results identify models and mechanisms of how negation, a compelling window into semantic representation, operates in real time.

Negation is ubiquitous – and therefore interesting in its own right. Furthermore, it offers a compelling linguistic framework to understand how the human brain builds meaning through combinatoric processes. Intuitively, negated concepts (e.g., "not good") entertain some relation with the affirmative concept (e.g., "good") as well as their counterpart (e.g., "bad"). The function of negation in natural language has been a matter of longstanding debate among philosophers, psychologists, logicians, and linguists [12]. In spite of its intellectual history and relevance (interpreting negation was, famously, a point of debate between Bertrand Russell and Ludwig Wittgenstein), comparatively little research investigates the cognitive and neural mechanisms underpinning negation. Previous work shows that negated phrases/sentences are processed with more difficulty (slower, with more errors) than the affirmative counterparts, suggesting an asymmetry between negated and affirmative representations; furthermore, state-of-the-art artificial neural networks appear to be largely insensitive to the contextual impacts of negation [13–20]. This asymmetry motivates one fundamental question: *how* does negation operate?

Studies addressing this question suggest that negation operates as a suppression mechanism by reducing the extent of available information [21–23], either in two steps [18,24–28] or in one incremental step [12,29–31]; other studies demonstrate that negation is rapidly and dynamically integrated into meaning representations [10,32], even unconsciously [33]. Within the context of action representation (e.g., "cut", "wish"), previous research suggests that negation recruits general-purpose inhibitory and cognitive control systems [34–41].

While the majority of neuroimaging studies focused on how negation affects action representation, psycholinguistic research shows that scalar adjectives (e.g., "bad-good", "close-open", "empty-full") offer insight into how negation operates on semantic representations of single words. These studies provide behavioral evidence that negation can either *eliminate* the negated concept and convey the opposite meaning ("not good" = "bad") or *mitigate* the meaning of its antonym along a semantic continuum ("not good" = "less good", "average", or "somehow bad"; [11,12,42–44]). Thus, the system of polar opposites generated by scalar adjectives provides an especially useful testbed to investigate changes in representation of abstract concepts along a semantic scale (e.g., "bad" to "good"), as a function of negation (e.g., "bad" vs. "not good").

Here, we capitalize on the semantic continuum offered by scalar adjectives to investigate *how* negation operates on the representation of abstract concepts (e.g., "bad" vs. "good"). First, we track how negation affects semantic representations over time in a behavioral study. Next, we use magnetoencephalography (MEG) and a decoding approach to track the evolution of neural representations of target adjectives in affirmative and negated phrases. We test four hypotheses: (1) negation does not change the representation of adjectives (e.g., "not good" = "good"), (2) negation weakens the representation of adjectives (e.g., "not good" < "good"), (3) negation inverts the representation of adjectives (e.g., "not good" = "bad"), and (4) negation changes the representation of adjectives to another representation (e.g., "not good" = e.g., "unacceptable"). The combined behavioral and neurophysiological data adjudicate among these hypotheses and identify potential mechanisms that underlie how negation functions in online meaning construction.

Results

Experiment 1: Continuous mouse tracking reveals a two-stage representation of negated adjectives

Experiment 1 (online behavioral experiment; N = 78) aimed to track changes in representation over time of scalar adjectives in affirmative and negated phrases. Participants read two-to-three-word phrases comprising one or two modifiers ("not" and "really") and a scalar adjective (e.g., "really really good", "really not quiet", "not ### fast"). The number and position of modifiers were manipulated to allow for a characterization of negation in simple and complex phrasal contexts, above and beyond single word processing. Adjectives were selected to represent opposite poles (i.e., antonyms) of the respective semantic scales: *low* pole of the scale (e.g., “bad”, “ugly”, “sad”, “cold”, “slow”, and “small”) and *high* pole of the scale (e.g., “good”, “beautiful”, “happy”, “hot”, “fast”, and “big”). A sequence of dashes was used to indicate the absence of a modifier. **Fig. 1A** and **Table S1** provide a comprehensive list of the linguistic stimuli. On every trial, participants rated the overall meaning of each phrase on a scale defined by each antonym pair (**Fig. 1A**). We analyzed reaction times and continuous mouse trajectories, which consist of the positions of the participant’s mouse cursor while rating the phrase meaning. Continuous mouse trajectories offer the opportunity to measure the unfolding of word and phrase comprehension over time, thus providing time-resolved dynamic data that reflect changes in meaning representation [15,45,46].

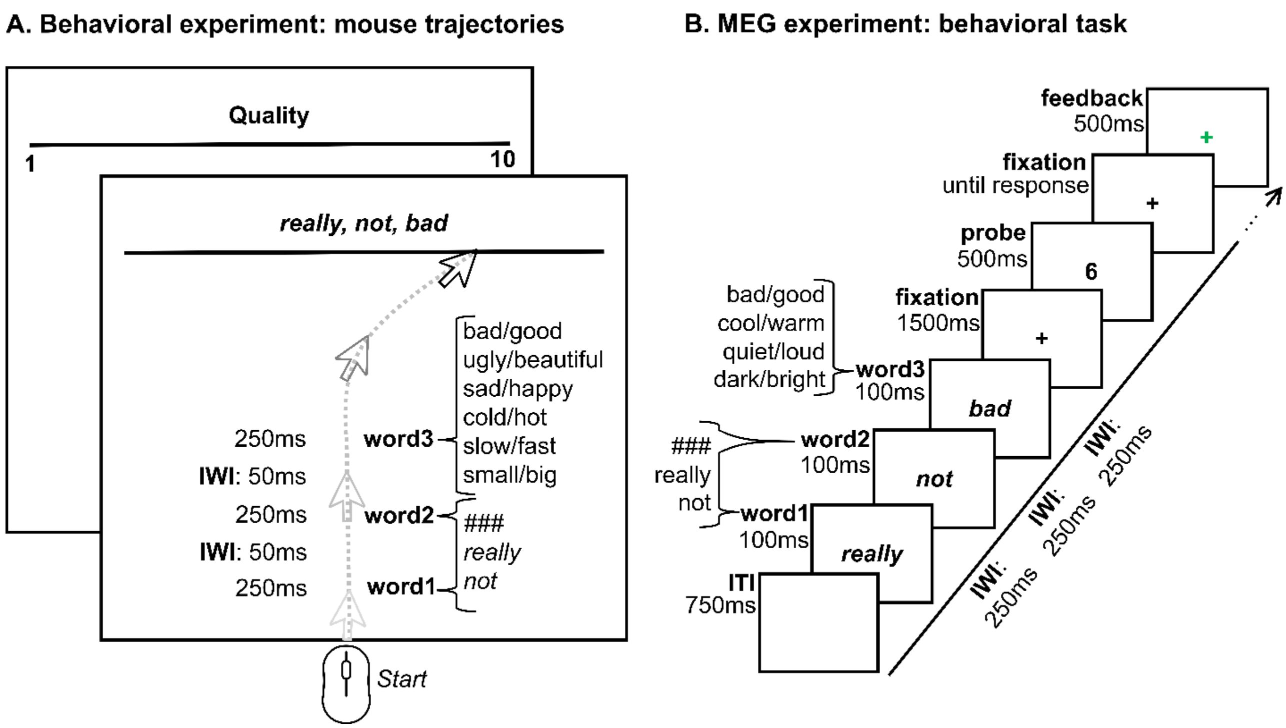


Figure 1. Experimental procedures.

(**A**) Behavioral procedure. Participants read affirmative or negated adjective phrases (e.g., “really really good”, “### not bad”) word by word and rated the overall meaning of each phrase on a scale. Each trial consisted of combinations of “###”, “really”, and “not” in word positions 1 and 2, followed by an adjective representing the low or high pole across six possible scalar dimensions. Before each trial, participants were informed about the scale direction, e.g., “bad” to “good”, i.e., 1 to 10. Scale direction was pseudorandomized across blocks. For each trial, we collected continuous mouse trajectories throughout the entire trial as well as reaction times. (**B**) MEG procedure. Participants read affirmative or negated adjective phrases and were instructed to derive the overall meaning of each adjective phrase on a scale from 0 to 8, e.g., from “really really bad” to “really really good”. After each phrase, a probe (e.g., 6) was presented, and participants were required to indicate whether the probe number correctly represented the overall meaning of the phrase on the scale (*yes/no* answer, using a keypad). Feedback was provided at the end of each trial (green or red cross). While performing the task, participants lay supine in a magnetically shielded room while continuous MEG data were recorded through a 157-channel whole-head axial gradiometer system. Panels A and B: "###" = no modifier; IWI = inter-word-interval.

Reaction times. To evaluate the effect of antonyms and of negation on reaction times in behavioral Experiment 1, we performed a 2 (*antonym*: low vs. high) x 2 (*negation*: negated vs. affirmative) repeated-measures ANOVA. The results revealed a significant main effect of antonyms (F(1,77) = 60.83, $p < 0.001$, $\eta_p^2 = 0.44$) and a significant main effect of negation (F(1,77) = 104.21, $p < 0.001$, $\eta_p^2 = 0.57$, **Fig.2A**). No significant crossover interaction between antonyms and negation was observed ($p > 0.05$). Participants were faster for high adjectives (e.g., "good") than for low adjectives (e.g., "bad") and for affirmative phrases (e.g., "really really good") than for negated phrases (e.g., "really not good"). These results support previous behavioral data showing that negation is associated with increased processing difficulty [15,16]. A further analysis including the number of modifiers as factor (i.e., *complexity*) indicates that participants were faster for phrases with two modifiers, e.g., "not really", than phrases with one modifier, e.g., "not ###" (F(1,77) = 16.02, $p < 0.001$, $\eta_p^2 = 0.17$), suggesting that the placeholder "###" may induce some interference to this otherwise relatively natural language task.

Continuous mouse trajectories. Continuous mouse trajectories across all adjective pairs and across all participants are depicted in **Fig.2B** and **Fig.2C** (*low* and *high* summarize the two antonyms across all scalar dimensions, see **Fig.S1** for each adjective dimension separately).

To quantify how the final interpretation of scalar adjectives changes as a function of negation, we first performed a 2 (*antonym*: low vs. high) x 2 (*negation*: negated vs. affirmative) repeated-measures ANOVA for participants' ends of trajectories (filled circles in **Fig.2B**), which reveal a significant main effect of antonyms (F(1,77) = 338.57, $p < 0.001$, $\eta_p^2 = 0.83$), a significant main effect of negation (F(1,77) = 65.50, $p < 0.001$, $\eta_p^2 = 0.46$), and a significant antonyms by negation interaction (F(1,77) = 1346.07, $p < 0.001$, $\eta_p^2 = 0.95$). Post-hoc tests show that the final interpretation of negated phrases is located at a more central portion on the semantic scale than that of affirmative phrases (affirmative low < negated high, and affirmative high > negated low, $p_{holm} < 0.001$). Furthermore, the final interpretation of negated phrases is significantly more variable (measured as standard deviations) than that of affirmative phrases (F(1,77) = 78.14, $p < 0.001$, $\eta_p^2 = 0.50$). Taken together, these results suggest that negation shifts the final interpretation of adjectives towards the antonyms, but never to a degree that overlaps with the interpretation of the affirmative antonym.

Second, we explored the temporal dynamics of adjective representation as a function of negation (i.e., from the presentation of word 1 to the final interpretation; lines in **Fig.2C**). While mouse trajectories of affirmative phrases branch towards either side of the scale and remain on that side until the final interpretation (lines in the left, gray, zoomed-in panel in **Fig.2C**), trajectories of negated phrases first deviate towards the side of the adjective and then towards the side of the antonym, to reach the final interpretation (i.e., "not low" first towards "low" and then towards "high"; right, gray, zoomed-in panel in **Fig.2C**; see **Fig.S1** for each adjective dimension separately). To characterize the degree of deviation towards each side of the scale, we performed regression analyses with antonyms as the predictor and mouse trajectories as the dependent variable (see **Methods**). The results confirm this observation, showing that (1) in affirmative phrases, betas are positive (i.e., mouse trajectories moving towards the adjective) starting at 300 ms from adjective onset ($p < 0.001$, green line in **Fig.2D**); and that (2) in negated phrases, betas are positive between 450 and 580 ms from adjective onset (i.e., mouse trajectories moving towards the adjective, $p = 0.04$), and only become negative (i.e., mouse trajectories moving towards the antonym, $p < 0.001$) from 700 ms from adjective onset (red line in **Fig.2D**). Note that beta values of negated phrases are smaller than that for affirmative phrases, again suggesting that negation does not invert the interpretation of the adjective to that of the antonym.

Finally, we replicated this experiment in a new group of 55 online participants (**Fig.S2**). The replication illustrates the robustness of the behavioral mouse tracking findings, even in the absence of feedback. Taken together, these results suggest that participants initially interpreted negated phrases as affirmative (e.g., "not good" interpreted along the "good" side of the scale) and later as a mitigated interpretation of the opposite meaning (e.g., the antonym "bad").

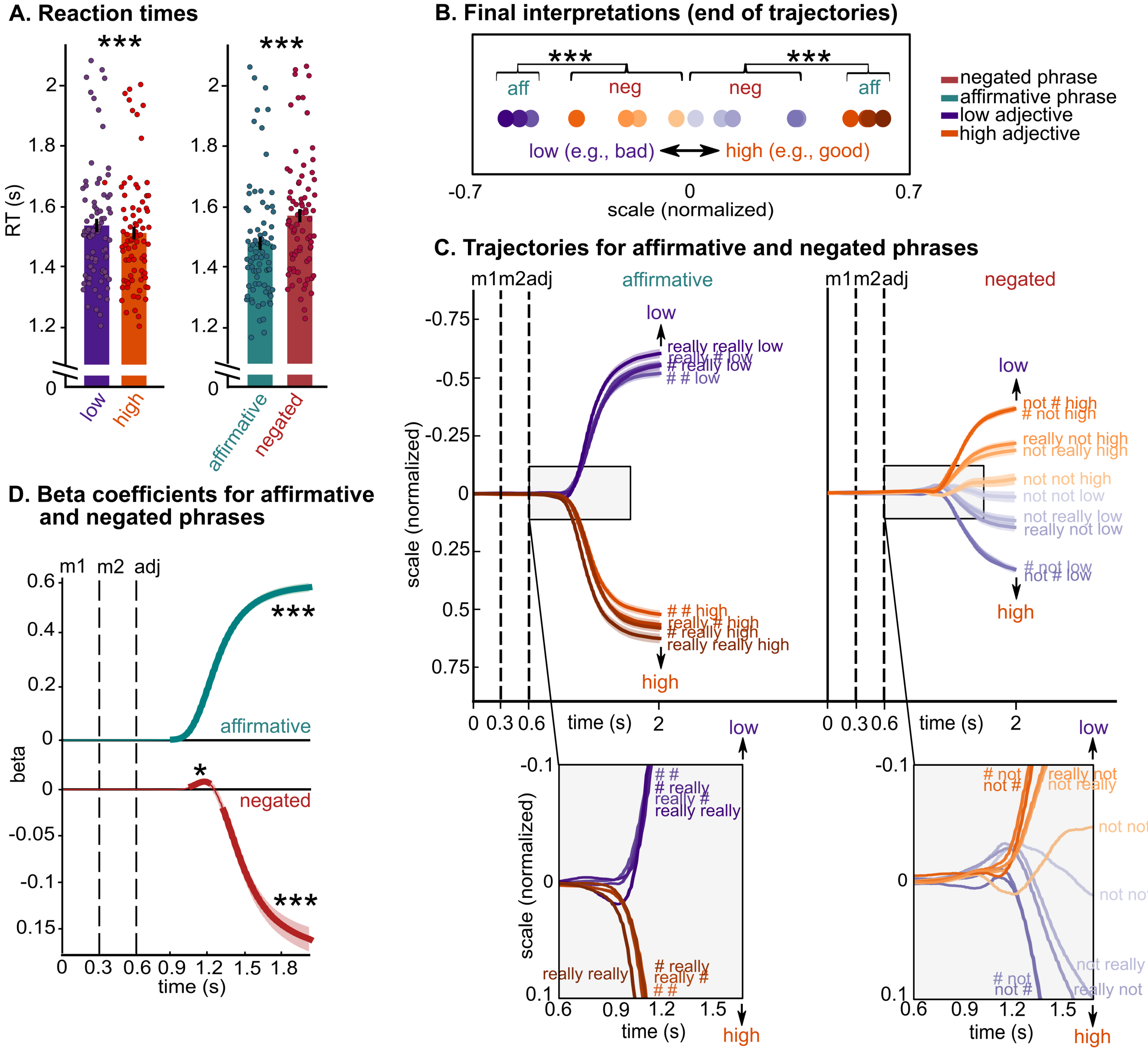


Figure 2. Behavioral results.

(A) Reaction times results for the online behavioral study (n=78). Bars represent the participants' mean ± SEM and dots represent individual participants. Participants were faster for high adjectives (e.g., "good") than for low adjectives (e.g., "bad") and for affirmative phrases (e.g., "really really good") than for negated phrases (e.g., "really not good"). The results support previous behavioral data showing that negation is associated with increased processing difficulty. **(B)** Final interpretations (i.e., end of trajectories) of each phrase, represented by filled circles (purple = low, orange = high), averaged across adjective dimensions and participants, showing that negation never inverts the interpretation of adjectives to that of their antonyms. **(C)** Mouse trajectories for low (purple) and high (orange) antonyms, for each modifier (shades of orange and purple) and for affirmative (left panel) and negated (right panel) phrases. Zoomed-in panels at the bottom demonstrate that mouse trajectories of affirmative phrases branch towards the adjective's side of the scale and remain on that side until the final interpretation; in contrast, the trajectories of negated phrases first deviate towards the side of the adjective and subsequently towards the side of the antonym. This result is confirmed by linear models fitted to the data at each timepoint in **D**. (**D**) Beta values (average over 78 participants) over time, separately for affirmative and negated phrases. Thicker lines indicate significant time windows. Panels **C**, **D**: black vertical dashed lines indicate the presentation onset of each word: modifier 1, modifier 2 and adjective; each line and shading represent participants' mean ± SEM; Panels A,B,D: *** $p < 0.001$; * $p < 0.05$.

Experiment 2: MEG shows that negation weakens the representation of adjectives and recruits response inhibition networks

In this study (MEG experiment, N = 26), participants read adjective phrases comprising one or two modifiers ("not" and "really") and scalar adjectives across different dimensions (e.g., "really really good", "really not quiet", "not ### dark"). Adjectives were selected to represent opposite poles (i.e., the antonyms) of the respective semantic scales: *low* pole of the scale (e.g., "bad", "cool", "quiet", "dark") and *high* pole of the scale (e.g., "good", "warm", "loud", "bright"). A sequence of dashes was used to indicate the absence of a modifier. **Fig. 1B** and **Table S2** provide the comprehensive list of the linguistic stimuli. Participants were asked to indicate whether a probe (e.g., 6) correctly represented the meaning of the phrase on a scale from "really really low" (0) to "really really high" (8) (*yes/no* answer, **Fig.1B**). Behavioral data of Experiment 2 replicate that of Experiment 1: negated phrases are processed slower and with more errors than affirmative phrases (main effect of negation for RTs: $F(1,25) = 26.44$, $p < 0.001$, $\eta_p^2 = 0.51$; main effect of negation for accuracy: $F(1,25) = 8.03$, $p = 0.009$, $\eta_p^2 = 0.24$).

The MEG analysis, using largely temporal and spatial decoding approaches [47], comprises four steps: (1) we first identify the temporal correlates of simple word representation (i.e., the words "really" and "not" in the modifier position, and each pair of scalar adjectives in the second word position, i.e., the head position); (2) we test lexical-semantic representations of adjectives over time beyond the single word level, by entering *low* ("bad", "cool", "quiet" and "dark") and *high* ("good", "warm", "loud" and "bright") antonyms in the same model. We then test the representation of the negation operator over time; (3) we then ask how negation operates on the representation of adjectives, by teasing apart four possible mechanisms (i.e., *No effect*, *Mitigation*, *Inversion*, *Change*); (4) we explore changes in beta power as a function of negation (motivated by the literature implicating beta-band neural activity).

(1) Temporal decoding of single word processing

Results show that the temporal decoding (see **Methods**) of "really" vs. "not" is significant between 120 and 430 ms and between 520 and 740 ms from the onset of the first modifier (dark gray areas, $p < 0.001$ and $p = 0.001$) and between 90 and 640 ms from the onset of the second modifier (light gray areas, $p < 0.001$, **Fig.3B**). Pairs of antonyms from different scales were similarly decodable between 90 and 410 ms from adjective onset (quality: 110 to 200 ms, $p = 0.002$ and 290 to 370 ms,

$p = 0.018$; temperature: 140 to 280 ms, $p < 0.001$; loudness: 110 to 410 ms, $p < 0.001$; brightness: 90 to 350 ms, $p < 0.001$, **Fig.3C**), reflecting time windows during which the brain represents visual, lexical, and semantic information (e.g., [7,48]).

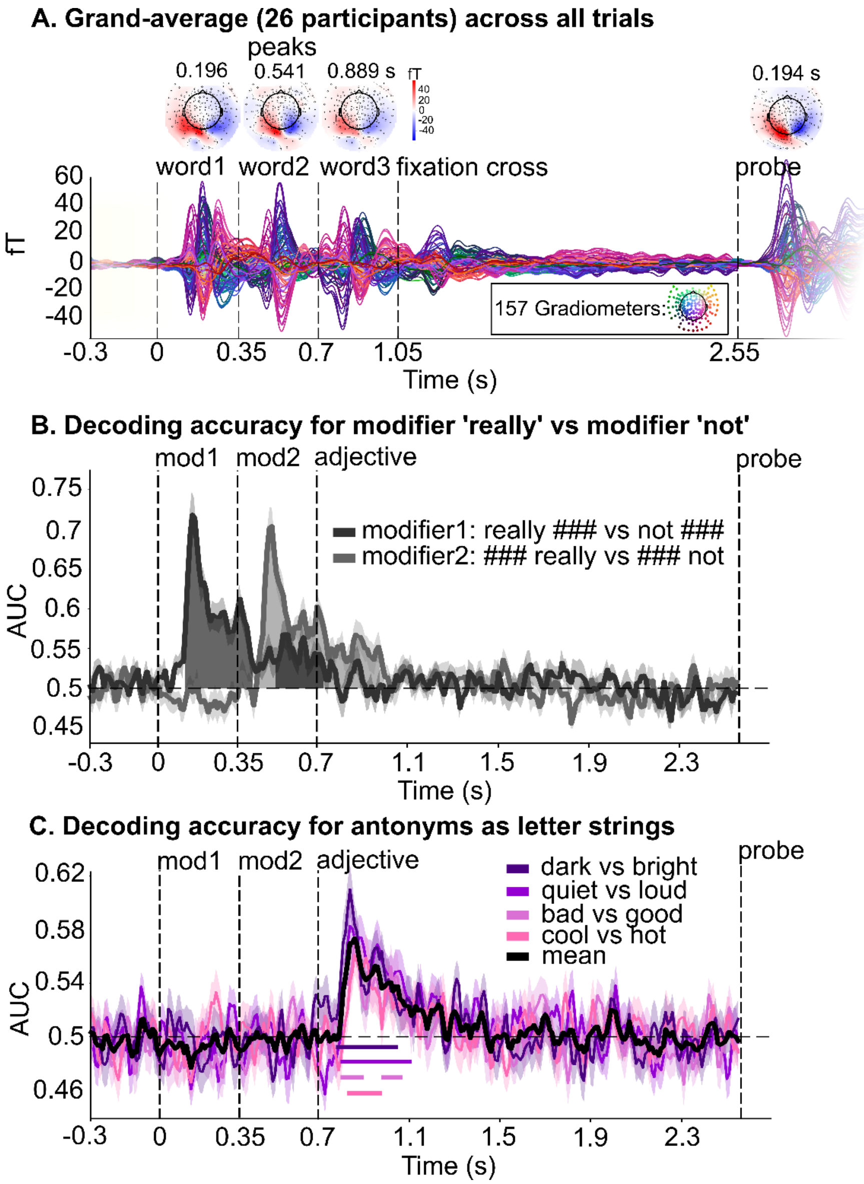


Figure 3. Evoked activity and temporal decoding of modifiers and adjectives as letter strings. **(A)** The butterfly (bottom) and topo plots (top) illustrate the event-related fields elicited by the presentation of each word as well as the probe, with a primarily visual distribution of neural activity right after visual onset (i.e., letter string processing). We performed multivariate decoding analyses on these preprocessed MEG data. Detector distribution of MEG system in inset box. fT: femtoTesla magnetic field strength. (**B)** We estimated the ability of the decoder to discriminate "really" vs. "not" in either modifier's position, from all MEG sensors. We contrasted phrases with modifiers "really ###" and "not ###", and phrases with modifiers "### not" and "### really". (**C)** We evaluated whether the brain encodes representational differences between each pair of antonyms (e.g., "bad" vs. "good"), in each of the four dimensions (quality, temperature, loudness, and brightness). The mean across adjective pairs is represented as a

solid black line; significant windows are indicated by horizontal solid lines below. For panels B and C: AUC = area under the receiver operating characteristic curve, chance = 0.5 (black horizontal dashed line); For all panels: black vertical dashed lines indicate the presentation onset of each word: modifier 1, modifier 2, and adjective; each line and shading represent participants' mean ± SEM.

(2) Temporal and spatial decoding of adjectives and negation

After establishing that single words' features can be successfully decoded in sensible time windows (see **Fig.3**), we moved beyond single word representation to selectively evaluate lexical-semantic differences between *low* ("bad", "cool", "quiet" and "dark") and *high* ("good", "warm", "loud" and "bright") adjectives, regardless of the specific scale (i.e., pooling over *quality*, *temperature*, *loudness*, and *brightness*). Temporal decoding analyses (see **Methods**) reveal significant decodability of *low* vs. *high* antonyms in three time windows between 140 and 560 ms from adjective onset (140 to 280 ms, $p < 0.001$; 370 to 460 ms: $p = 0.009$; 500 to 560 ms: $p = 0.044$, purple areas in **Fig.4A**). No significant differences in lexical-semantic representation between *low* and *high* antonyms were observed in later time windows (i.e., after 560 ms from adjective onset). The spatial decoding analysis illustrated in **Fig.4B** (limited to 50-650 ms from adjective onset, see **Methods**) show that decoding accuracy for *low* vs. *high* antonyms is significantly above chance in a widespread left-lateralized brain network, encompassing the anterior portion of the superior temporal lobe, the middle, and the inferior temporal lobe (purple areas in **Fig.4B**, significant clusters are indicated by a black contour: left temporal lobe cluster, $p = 0.002$). A significant cluster was also found in the right temporal pole, into the insula ($p = 0.007$). Moreover, we found significant clusters in the bilateral cingulate gyri (posterior and isthmus) and precunei (left precuneus/cingulate cluster, $p = 0.009$; right precuneus/cingulate cluster, $p = 0.037$). Overall, these regions are part of the (predominantly left-lateralized) frontotemporal brain network that underpins lexical-semantic representation and composition [7,8,48–55].

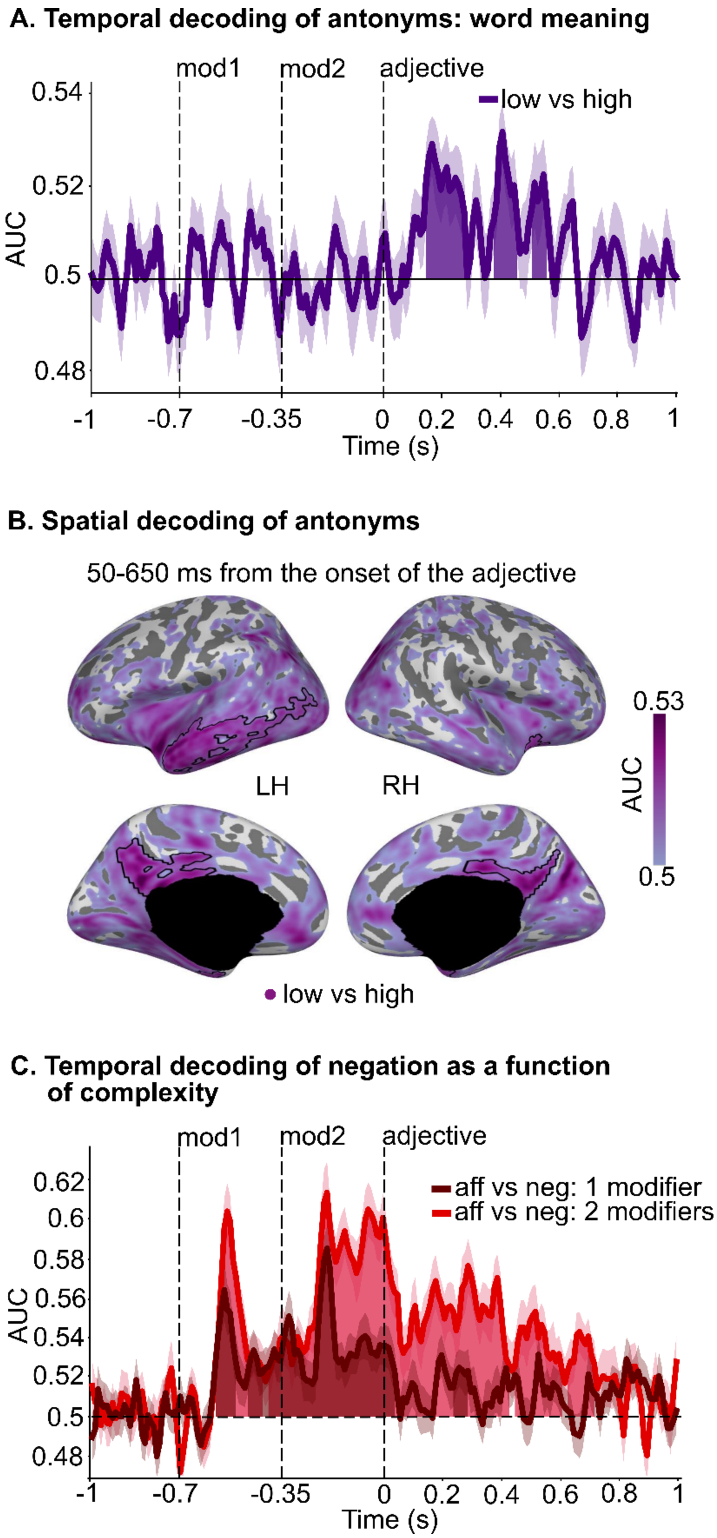


Figure 4. Temporal and spatial decoding of antonyms across all scales and temporal decoding of negation. (**A**) Decoding accuracy (purple line) of lexical-semantic differences between antonyms across all scales (i.e., pooling over "bad", "cool", "quiet" and "dark"; and "good", "warm", "loud" and "bright" before fitting the estimators) over time; significant time windows are indicated by purple areas; (**B**) Decoding accuracy (shades of purple) for antonyms across all scales over brain sources (after pooling over the four dimensions), between 50 and 650 ms from adjective onset. Significant spatial clusters are indicated by a black contour. (**C**) Decoding accuracy of negation over time, as a

function of the number of modifiers (1 modifier: dark red line and shading; 2 modifiers: light red line and shading). Significant time windows are indicated by dark red (1 modifier) and light red (2 modifiers) areas. For all panels: AUC: area under the receiver operating characteristic curve, chance = 0.5 (black horizontal dashed line); black vertical dashed lines indicate the presentation onset of each word: modifier1, modifier2 and adjective; each line and shading represent participants' mean ± SEM; aff = affirmative, neg = negated; LH = left hemisphere; RH = right hemisphere.

Next, we turn to representations of negation over time. We performed a temporal decoding analysis for phrases containing "not" vs. phrases not containing "not", separately for phrases with one and two modifiers (to account for phrase complexity; see **Table S2** for a list of all trials). For phrases with one modifier, the decoding of negation is significantly higher than chance throughout word 1 (-580 to -500 ms from adjective onset, $p = 0.005$), then again throughout word 2 (-470 to 0 ms from adjective onset, $p < 0.001$). After the presentation of the adjective, negation decodability is again significantly above chance between 0 and 40 ms ($p = 0.034$) and between 230 and 290 ms from adjective onset ($p = 0.018$; dark red line and shading in **Fig.4C**). Similarly, for phrases with two modifiers, the decoding of negation is significantly higher than chance throughout word 1 (-580 to -410 ms from adjective onset, $p = 0.002$), throughout word 2 (-400 to 0 ms from adjective onset, $p < 0.001$), and for a longer time window from adjective onset compared to phrases with one modifier, i.e., between 0 and 720 ms (0 to 430 ms, $p < 0.001$; 440 to 500 ms, $p = 0.030$; 500 to 610 ms, $p < 0.001$; 620 to 720 ms, $p < 0.001$; light red line and shading in **Fig.4C**). The same analysis time-locked to the onset of the probe shows that negation is once again significantly decodable between 230 and 930 ms after the probe (**Fig.S3**).

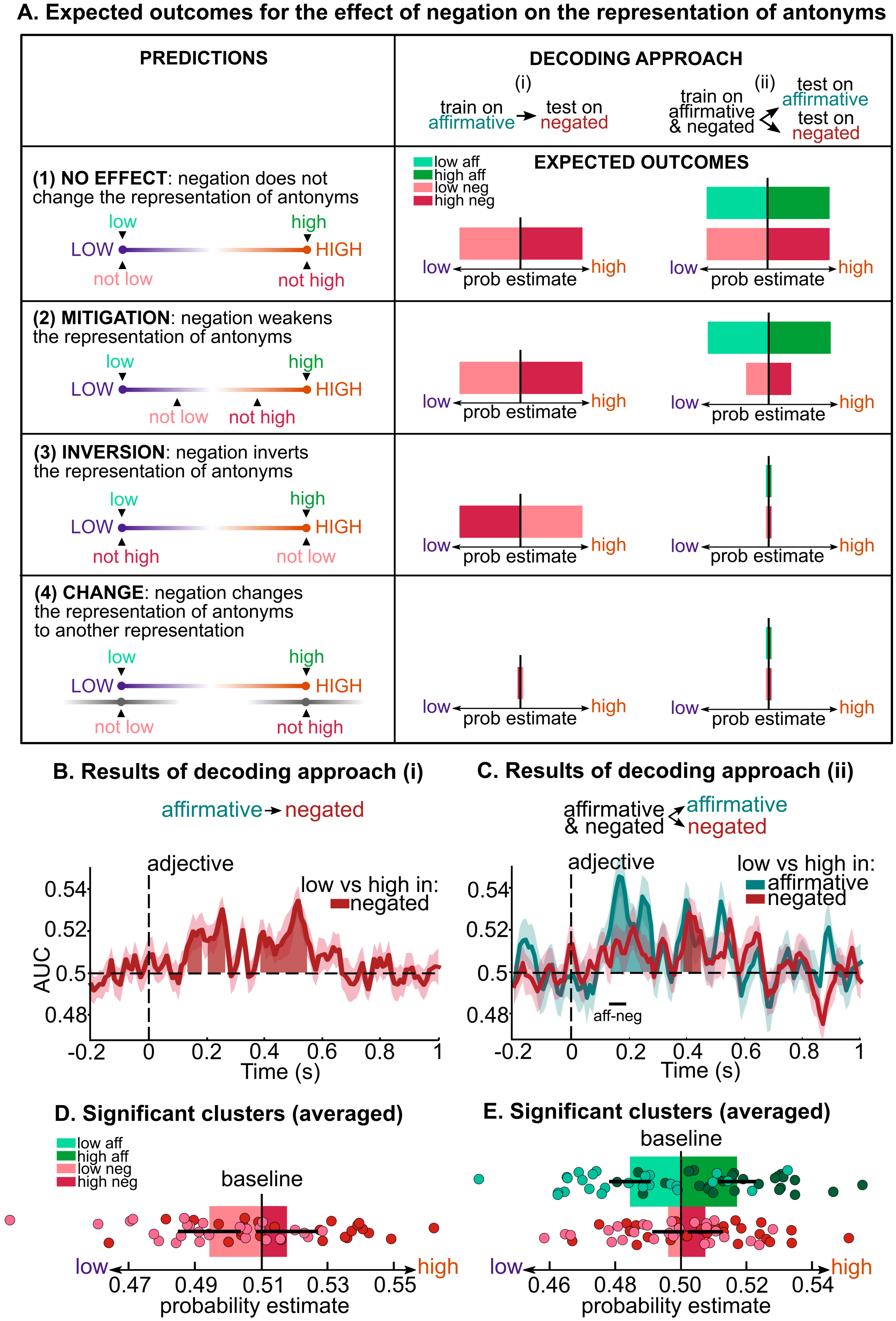
Cumulatively, these results suggest that the brain encodes negation every time a "not" is presented and maintains this information up to 720 ms after adjective onset. Further, they show that the duration of negation maintenance is amplified by the presence of a second modifier, highlighting combinatoric effects [2,6,56].

(3) Effect of negation on lexical-semantic representations of antonyms over time

The temporal decoding analyses performed separately for adjectives and for negation demonstrates that the brain maintains the representation of the modifiers available throughout the presentation of the adjective. Here we ask how negation *operates on* the representation of the antonyms at the neural level, leveraging theoretical accounts of negation [11,12,42–44], behavioral results of Experiment 1, and two complementary decoding approaches. We test four hypotheses (see *Predictions* in **Fig.5A**): (1) *No effect of negation*: negation does not change the representation of adjectives (i.e., "not low" = "low"). We included this hypothesis based on the two-step theory of negation, wherein the initial representation of negated adjectives would not be affected by negation [27]. (2) *Mitigation*:

negation weakens the representation of adjectives (i.e., "not low" < "low"). (3) *Inversion*: negation inverts the representation of adjectives (i.e., "not low" = "high"). Hypotheses (3) and (4) are derived from previous linguistics and psycholinguistics accounts on comprehension of negated adjectives [42–44]. Finally, (4) *Change*: we evaluated the possibility that negation might change the representation of adjectives to another representation outside the semantic scale defined by the two antonyms (e.g., "not low" = e.g., "fair").

To adjudicate between these four hypotheses, we performed two sets of decoding analyses. Decoding approach (i): we computed the accuracy with which estimators trained on *low* vs. *high* antonyms in affirmative phrases (e.g., "really really bad" vs. "really really good") generalize to the representation of *low* vs. *high* antonyms in negated phrases (e.g., "really not bad" vs. "really not good") at each time sample time-locked to adjective onset (see **Methods**); decoding approach (ii): we trained estimators on *low* vs. *high* antonyms in affirmative and negated phrases together (in 90% of the trials) and computed the accuracy of the model in predicting the representation of *low* vs. *high* antonyms in affirmative and negated phrases separately (in the remaining 10% of the trials; see **Methods**). Decoding approach (ii) allows for direct comparison between AUC and probability estimates in affirmative and negated phrases. Expected probability estimates (i.e., the averaged class probabilities for *low* and *high* classes) as a result of decoding approach (i) and (ii) are depicted as light and dark, green and red bars under *Decoding approach* in **Fig.5A**.



360 **Figure 5. Predictions, decoding approaches, and results of the effect of negation on the representation of**

361 **adjectives.**

362 (**A**) We tested four possible effects of negation on the representation of adjectives: (1) *No effect*, (2) *Mitigation*, (3)

363 *Inversion*, (4) *Change* (left column). Note that we depicted predictions of (3) *Inversion* on the extremes of the scale,

364 but a combination of inversion and mitigation would predict the same outcomes. We performed two sets of decoding

analyses (right column): (i) We trained estimators on low (purple) vs. high (orange) antonyms in affirmative phrases and predicted model accuracy and probability estimates of low vs. high antonyms in negated phrases (light and dark red bars). (ii) We trained estimators on low vs. high antonyms in affirmative and negated phrases together and predicted model accuracy and probability estimates in affirmative (light and dark green bars) and negated phrases (light and dark red bars) separately. (**B**) Decoding accuracy (red line) over time of antonyms for negated phrases, as a result of decoding approach (i). Significant time windows are indicated by red areas. (**C**) Decoding accuracy of antonyms over time for affirmative (green line) and negated (red line) phrases, as a result of decoding approach (ii). Significant time windows for affirmative and negated phrases are indicated by green and red areas. The significant time window of the difference between affirmative and negated phrases is indicated by a black horizontal solid line. (**D**) Probability estimates for low (light red) and high (dark red) negated antonyms averaged across the significant time windows depicted in **B**. Bars represent the participants' mean ± SEM and dots represent individual participants. (**E**) Probability estimates for low (light green) and high (dark green) affirmative adjectives and for low (light red) and high (dark red) negated adjectives, averaged across the significant time window depicted as a black horizontal line in **C.** Chance level of probability estimates was computed by averaging probability estimates of the respective baseline (note that the baseline differs from 0.5 due to the different number of trials for each class in the training set of decoding approach (i)). Bars represent the participants' mean ± SEM and dots represent individual participants. For panels **B** and **C**: AUC: area under the receiver operating characteristic curve, chance = 0.5 (black horizontal dashed line); each line and shading represent participants' mean ± SEM. Panels B,C,D,E: the black vertical dashed line indicates the presentation onset of the adjective; green = affirmative phrases, red = negated phrases.

Temporal decoding approach (i) reveals that the estimators trained on the representation of *low* vs. *high* antonyms in affirmative phrases significantly generalize to the representation of *low* vs. *high* antonyms in negated phrases, in four time windows between 130 and 550 ms from adjective onset (130 to 190 ms, $p = 0.039$; 200 to 270 ms: $p = 0.003$; 380 to 500 ms: $p < 0.001$; 500 to 550 ms: $p = 0.008$; red areas in **Fig.5B**). **Fig.5D** depicts the probability estimates averaged over the significant time windows for *low* and *high* antonyms in negated phrases. These results only support predictions (1) *No effect* and (2) *Mitigation*, thus invalidating predictions (3) *Inversion* and (4) *Change*. **Fig.S4** illustrates a different approach that similarly leads to the exclusion of prediction *(3) Inversion*.

Temporal decoding approach (ii) shows significant above chance decoding accuracy for affirmative phrases between 130 and 280 ms ($p < 0.001$) and between 370 and 420 ms ($p = 0.035$) from adjective onset. Conversely, decoding accuracy for negated phrases is significantly above chance only between 380 and 450 ms after the onset of the adjective ($p = 0.004$). Strikingly, negated phrases are associated with significantly lower decoding accuracy than affirmative phrases in the time window between 130 and 190 ms from adjective onset ($p = 0.040$; black horizontal line in **Fig.5C**). **Fig.5E** represents the probability estimates averaged over this 130-190 ms significant time window for *low* and *high* antonyms, separately in affirmative and negated phrases, illustrating reduced probability estimates for negated compared to affirmative phrases.

Overall, the generalization of representation from affirmative to negated phrases and the higher decoding accuracy (and probability estimates) for affirmative than negated phrases within the first 500 ms from adjective onset (i.e., within the time window of lexical-semantic processing shown in **Fig.4A**) provide direct evidence in support of prediction (2) *Mitigation*, wherein negation weakens the representation of adjectives. The alternative hypotheses did not survive the different decoding approaches.

(4) Changes in beta power as a function of negation

We distinguished among four possible mechanisms of how negation could operate on the representation of adjectives and demonstrated that negation does not invert or change the representation of adjectives but rather weakens the decodability of *low* vs. *high* antonyms, significantly for about 60 ms from adjective onset. The availability of negation upon the processing of the adjective (**Fig.4C**) and the reduced decoding accuracy for antonyms in negated phrases (**Fig.5C**) raise the question of whether negation operates through inhibitory mechanisms, as suggested by previous research employing action-related verbal material [35–37]. We therefore performed time-frequency analyses, focusing on beta power (including low-beta: 12 to 20 Hz, and high-beta: 20 to 30 Hz, [57], see **Methods**), which has been previously associated with inhibitory control [58] (see **Fig.S5** for comprehensive time-frequency results). We reasoned that, if negation operates through general-purpose inhibitory systems, we should observe higher beta power for negated than affirmative phrases in sensorimotor brain regions.

Our results are consistent with this hypothesis, showing significantly higher low-beta power (from 229 to 350 ms from the onset of modifier1: $p = 0.036$; from 326 to 690 ms from adjective onset: $p = 0.012$; red line in **Fig.6A**) and high-beta power (from 98 to 271 ms from adjective onset: $p = 0.044$; yellow line in **Fig.6A**) for negated than affirmative phrases. **Fig.S6** further shows low and high-beta power separately for negated and affirmative phrases, compared to phrases with no modifier.

Our whole-brain source localization analysis shows significantly higher low-beta power for negated than affirmative phrases in the left precentral, postcentral, and paracentral gyri ($p = 0.012$; between 326 and 690 ms from adjective onset, red cluster in **Fig.6C**). For high-beta power, similar (albeit not significant) sensorimotor spatial patterns emerge (yellow cluster in **Fig.6B**).

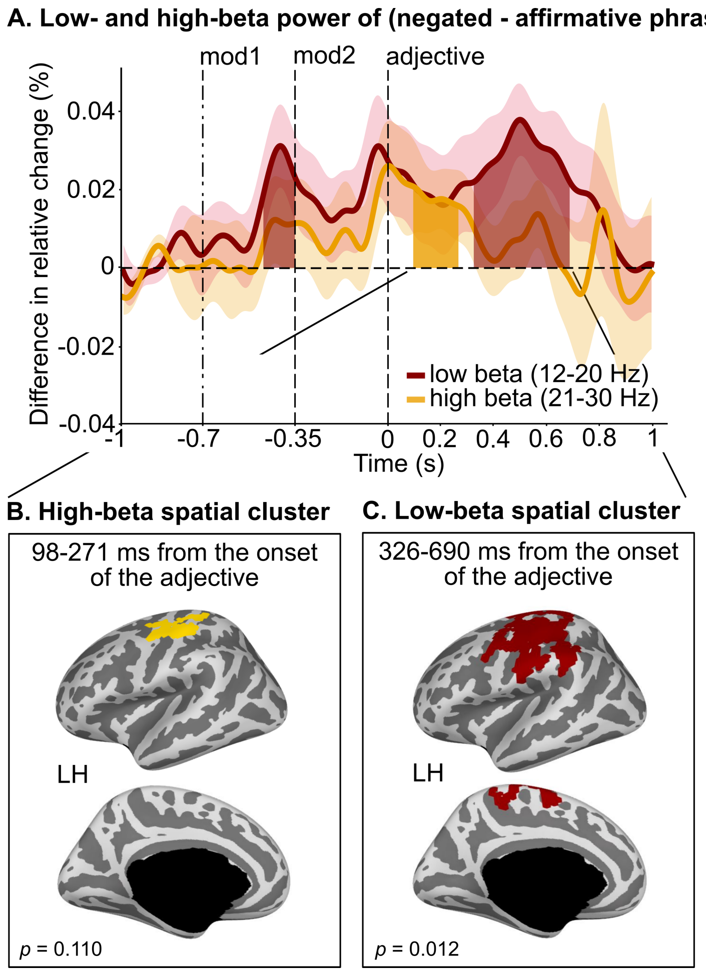


Figure 6. Differences in beta power over time between negated and affirmative phrases.
(**A**) Differences in low (12-20 Hz, red) and high (21-30 Hz, yellow) beta power over time between negated and affirmative phrases. Negated phrases show higher beta power compared to affirmative phrases throughout the presentation of the modifiers and for a sustained time window from adjective onset up to ~700 ms; significant time windows are indicated by red (low-beta) and yellow (high-beta) areas; black vertical dashed lines indicate the presentation onset of each word: modifier1, modifier2 and adjective; each line and shading represent participants' mean ± SEM. (**B**) Differences (however not reaching statistical significance, α = 0.05) in high-beta power between negated and affirmative phrases (restricted between 97 and 271 ms from adjective onset, yellow cluster). (**C**) Significant differences in low-beta power between negated and affirmative phrases (restricted between 326 and 690 ms from adjective onset) in the left precentral, postcentral and paracentral gyrus (red cluster). Note that no significant spatial clusters were found in the right hemisphere.

Discussion

We tracked changes over time in lexical-semantic representations of scalar adjectives, as a function of the intensifier "really" and the negation operator "not". Neural correlates of negation have typically been investigated in the context of action verbs [29,35–37,40,41,59–63]. Our study employs minimal linguistic contexts to characterize in detail how negation operates on abstract, non-action-related lexical-semantic representations. We leveraged (1) psycholinguistic findings on adjectives that offer a framework wherein meaning is represented on a continuum [42,43], (2) time-resolved behavioral and neural data, and (3) multivariate analysis methods (decoding) which can discriminate complex lexical-semantic representations from distributed neuronal patterns (e.g., [62]).

The longer RTs and decreased accuracy for negated phases shown in Experiment 1 (**Fig.2A**), in the replication experiment (**Fig.S2A**), and in Experiment 2, are consistent with data demonstrating that negation incurs increased processing costs [13–18,27,32]. More significantly, mouse trajectories show that participants initially interpreted negated phrases as affirmative (e.g., "not good" is located on the "good" side of the scale, for ~130 ms, **Fig.2C** and **Fig.S2C**), indicating that initial representations of negated scalar adjectives are closer to the representations of the adjectives rather than that of their antonyms. Similarly, participants' final interpretations of negated adjectives (e.g., "not good", "really not good") never overlapped with the final interpretations of the corresponding affirmative antonyms (e.g., "bad", "really bad", "really really bad"; **Fig.2B** and **Fig.S2B**) highlighting how negation never inverts the meaning of an adjective to that of its antonym, even when participants are making decisions on a binary semantic scale (9,37-40).

Continuous mouse trajectories allowed us to quantify dynamic changes in participants' interpretations. MEG provided a means to directly track neural representations over time. We first identified the temporal correlates of lexical-semantic processing *separately* for scalar adjectives and for the negation operator. The time window of adjective representation (~140-560 ms from adjective onset, **Fig.4A**) is consistent with previous studies investigating lexical-semantic processing in language comprehension (130–200 ms up to ~550 ms from word onset [64–68]). Spatial decoding results corroborate temporal results, highlighting the involvement of the left-lateralized frontotemporal brain network in adjective processing (**Fig.4B,** [7,8,48–55]). Our data further demonstrate that negation is processed in parallel to the processing of the adjective (up to ~700 ms; **Fig.4C**), not serially (see [69,70] for related patterns in the context of negation + auxiliary verb and adjective + noun). Finally, they show that the decodability of negation increases in phrases with two modifiers (e.g., "really not", "not really", **Fig.4C, Fig.S3**), highlighting compositional effects [6].

We then evaluated the effects of the negation operator *on* adjective representation, to address the question of *how* negation operates on lexical-semantic representations of antonyms. We contrasted four hypotheses (**Fig.5A**): negation (1) does not change the representation of scalar adjectives (e.g., "not good" = "good", *No effect*), (2) weakens the representation of scalar adjectives (e.g., "not good" < "good", *Mitigation*), (3) inverts the representation of scalar adjectives (e.g., "not good" = "bad", *Inversion*), or (4) changes the representation of scalar adjectives to another representation (e.g., "not good" = e.g., "unacceptable", *Change*). First, we demonstrated that, within the time window of adjective encoding, the representation of affirmative adjectives generalizes to that of negated adjectives (**Fig.5B** and **Fig.5D**). This finding rules out predictions (3) *Inversion* and (4) *Change*. Moreover, these findings complement our behavioral data that show that negated adjectives are initially interpreted by participants as affirmative. Second, we showed that the representation of adjectives in affirmative and negated phrases is not identical but is weakened by negation (**Fig.5C** and **Fig.5E**). This result rules out prediction (1) *No effect* and supports prediction (2) *Mitigation*, wherein negation weakens the representation of adjectives. We observed such reduction in early lexical-semantic representations (i.e., from ~130 ms post adjective-onset), supporting previous research that reported effects of negation as soon as lexical-semantic representations of words are formed [12,29–31,71], and not exclusively at later processing stages (e.g., P600 [72,73]).

Our behavioral and neural data jointly point to a *mitigation* rather than an *inversion* effect of negation: initial interpretations and neural representations of negated adjectives are similar to that of affirmative adjectives, but weakened; final interpretations do not overlap with neither affirmative extreme of the semantic scale. While previous fMRI studies on sentential negation have shown that negation reduces hemodynamic brain activations related to verb processing [40,41], the current study offers novel time-resolved behavioral and neural data on how negation selectively operates on abstract concepts. Previous research has highlighted that negation might behave differently depending on the pragmatics of discourse interpretation, e.g., when presented in isolation as compared to when presented in context ("not wrong" vs. "your theory is not wrong" [9,10]), or when used ironically ("they are not really good" said ironically to mean that they are "mediocre", e.g., [11,71]). Within this pragmatic framework, it has been suggested that the opposite meaning of a scalar adjective would be more simply conveyed by the affirmative counterpart than by negation [11,44,74]; thus, to convey the opposite meaning of "bad", it would be more appropriate to use "good" as opposed to "not bad". Following this logic, negation would be purposefully used (and understood) to convey a different, mitigated meaning of the adjective (e.g., "not bad" = "less than bad"). Although we did not directly manipulate sentential or pragmatic contexts, our findings provide behavioral and neural evidence that negation acts as a mitigator. Here we only tested adjective pairs that form *contraries* (which lie on a continuum, e.g., "bad" and "good"); thus inherently different patterns of results could emerge in the case of *contradictories* (which form a dichotomy, e.g., "dead" and "alive", [44]), where there is no continuum for mitigation to have an effect.

Overall, evidence that negation weakens adjective representations invites the hypothesis that negation operates as a suppression mechanism, possibly through general-purpose inhibitory systems [36,37]. To address this, we compared beta power modulations in affirmative and negated phrases (**Fig.6**). In addition to subserving motor processing, beta-power modulation (12-30 Hz) is associated with multiple aspects of language processing [35,75–78]; for a review, see [57,79]). We evaluated differences between negated and affirmative phrases separately in the low- and high-beta bands. We found greater power for negated than affirmative phrases in both bands, during the processing of the modifier and throughout the processing of the adjective up to ~700 ms, localized in left-lateralized sensorimotor areas. The timing and spatial correlates of beta-power in relation to negation align with studies that examined the effect of negation on (mental and motor) action representation [36]. Strikingly, we demonstrated that negation recruits brain areas and neurophysiological mechanisms similar to that recruited by response inhibition - however in the absence of action-related language material. Within a framework that recognizes two interactive neural systems, i.e., a semantic representation and a semantic control system [53], negation would operate through the latter, modulating how activation propagates through the (ventral) language semantic network wherein meaning is represented. The precise connectivity that underpins mitigation of lexical-semantic representations remains to be investigated.

Collectively, we demonstrated that, by characterizing subtle changes of linguistic meaning through negation, using time-resolved behavioral and neuroimaging methods and multivariate decoding, we can tease apart different possible representation outcomes of combinatorial operations, above and beyond the sum of the processing of individual word meanings.

Materials and Methods

Participants

Experiment 1: continuous behavioral tracking. 101 participants (46 females; mean age = 29.6 years; range 18-67 years) completed an online mouse tracking experiment. Participants were recruited via Amazon Mechanical Turk and via the platform SONA (a platform for students' recruitment). All participants were native English speakers with self-reported normal hearing, normal or corrected to normal vision, and no neurological deficits. 97 participants were right-handed. Participants were paid or granted university credits for taking part in the study, which was performed online. All participants provided written informed consent, as approved by the local institutional review board (New York University's Committee on Activities Involving Human Subjects). The data of 23 participants were excluded from the data analysis due to (i) number of "incorrect" feedback (based on the warnings) > 30%, (ii) mean RTs > 2SD from the group mean, or (iii) response trajectory always ending within 1/4 from the center of the scale, regardless of condition (i.e., participants who did not pay attention to the instructions of the task). Thus, 78 participants were included in the analysis. The sample size was determined based on previous studies using a similar behavioral approach (~30 participants [15,45,80]) and was increased to account for the exclusion rate reported for online crowdsourcing experiments [81,82]. For participants in *Experiment 1 (replication)* see **Fig.S2**.

Experiment 2: MEG. A new group of 28 participants (17 females; mean age = 28.7 years; range 19-53 years) took part in the in-lab MEG experiment. All participants were native English speakers with self-reported normal hearing, normal or corrected to normal vision, and no neurological deficits. 24 participants were right-handed. They were paid or granted university credits for taking part in the study. All participants provided written informed consent, as approved by the local institutional review board (New York University's Committee on Activities Involving Human Subjects). The data of 2 participants were excluded from the data analysis because their accuracy scores in the behavioral task was < 60%. Thus, 26 participants were included in the analysis. The sample size was determined based on previous studies investigating negation using EEG (17 to 33 participants [26,35,37]), investigating semantic representation using MEG (25 to 27 participants [7,8]), or employing decoding methods with MEG data (17 to 20 participants [83,84]).

Stimuli, Design, and Procedure

Experiment 1 (and replication): continuous mouse tracking.

Stimuli and Design. The linguistic stimulus set comprises 108 unique adjective phrases (for the complete list, see **Table S1**). Adjectives were selected to be antonyms (i.e., *low* and *high* poles of the scale) in the following six cognitive or sensory dimensions: *quality* ("bad", "good"), *beauty* ("ugly", "beautiful"), *mood* ("sad", "happy"), *temperature* ("cold", "hot"), *speed* ("slow", "fast"), and *size* ("small", "big"). These antonyms are all *contraries* (i.e., adjectives that lie on a continuum [44]). Lexical characteristics of the antonyms were balanced according to the English Lexicon Project [85]; mean (SD) HAL log frequency of *low* adjectives: 10.69 (1.09), *high* adjectives: 11.51 (1.07), mean (SD) bigram frequency of *low* adjectives: 1087.10 (374), *high* adjectives: 1032 (477.2); mean

(SD) lexical decision RTs of *low* adjectives: 566 (37), *high* adjectives: 586 ms (70)). Adjectives were combined with zero (e.g., "### ###"), one (e.g., "really ###"), or two modifiers (e.g., "really not"). Modifiers were either the intensifier "really" or the negation "not" (see [33] for a similar choice of modifiers). A sequence of dashes was used to indicate the absence of a modifier, e.g., "really ### good". Each of the 12 adjectives was preceded by each of the nine possible combinations of modifiers: "### ###", "### really", "really ###", "### not", "not ###", "really not", "not really", "really really" and "not not" ("not not" was included to achieve a full experimental design, even if it is not a frequent combination in natural language and its cognitive and linguistic representations are still under investigation[86]). Each dimension (e.g., quality) was presented in two blocks (one block for each scale orientation, e.g., low to high and high to low) for a total of 12 blocks. Each phrase was repeated three times within each block (note that "### really"/"really ###" were repeated an overall of three times, and so were "### not"/"not ###"). Thus, the overall experiment comprised 504 trials. The order of phrases was randomized within each block for each participant. The order of pairs of blocks was randomized across participants.

Procedure. Behavioral trajectories provide time-resolved dynamic data that reflect changes in representation [15,45,46]. The online experiment was developed using oTree, a Python-based framework for the development of controlled experiments on online platforms [87]. Participants performed this study remotely, using their own monitor and mouse (touchpads were not allowed). They were instructed to read affirmative or negated adjective phrases (e.g., "really really good", "really not bad") and rate the overall meaning of each phrase on a scale, e.g., from "really really bad" to "really really good". Participants were initially familiarized with the experiment through short videos and a short practice block (18 trials with feedback). They were instructed that the poles of the scale (e.g., "bad" and "good") would be reversed in half of the trials and warned that (i) they could not cross the vertical borders of the response space, (ii) they had to maintain a constant velocity, by following an horizontal line moving vertically, and (iii) they could not rate the meaning of the phrase before the third word was presented. At the beginning of each trial, a response area of 600 (horizontal) x 450 (vertical) pixels and a solid line at the top of the rectangle were presented (**Fig.1A**). Participants were informed about the scale (e.g., quality) and the direction of the scale (e.g., "bad" to "good" or "good" to "bad", i.e., 1 to 10 or 10 to 1). Participants were instructed to click on the "start" button and move the cursor of the mouse to the portion of the scale that best represented the overall meaning of the phrase. The "start" button was placed in the center portion of the bottom of the response space (i.e., in a neutral position). Once "start" was clicked on, information about the scale and scale direction disappeared, leaving only the solid line on screen.

Phrases were presented at the top of the response space, from the time when participants clicked on "start", one word at a time, each word for 250 ms (inter-word-interval: 50 ms). After each trial, participants were provided the "incorrect" feedback if the cursor's movement violated the warnings provided during the familiarization phase, and an explanation was provided (e.g., "you crossed the vertical borders"). To keep participants engaged, we provided feedback also based on the final interpretation: "incorrect" if the response was in the half of the scale opposite to the adjective (for the conditions: "### ###", "#### really", "really ###" and "really really"), or in the same half of the scale of the adjective (for the conditions: "### not" or "not ###"), or in the outer 20% left and right portions of the scale (for the conditions: "really not", "not really" and "not not"); feedback was "correct" otherwise. In case of an "incorrect" trial, the following trial was delayed for 4 seconds. For each trial, we collected continuous mouse trajectories and RTs. The overall duration of the behavioral experiment was approximately 90 minutes. To verify that the feedback did not affect our results, we ran a replication study with 55 online participants where no feedback was provided based on the final interpretation (**Fig.S2**).

Experiment 2: MEG.

Stimuli and Design. The linguistic stimulus set comprised 72 unique adjective phrases (for the complete list, see **Table S2**). Similar to the Experiment 1, adjectives were selected for being antonyms (and *contraries*) in the following cognitive or sensory dimensions: *quality* ("bad", "good"), *temperature* ("cool", "warm"), *loudness* ("quiet", "loud"), and *brightness* ("dark", "bright"). Lexical characteristics of the antonyms were balanced according to the English Lexicon Project ([85]; mean (SD) HAL log frequency of "low" adjectives: 10.85 (1.03), "high" adjectives: 10.55 (1.88); mean (SD) bigram frequency of "low" adjectives: 1196.5 (824.6), "high" adjectives: 1077.5 (376.3); mean (SD) lexical decision RTs of "low" adjectives: 594 ms (39), "high" adjectives: 594 (33)). Adjectives were combined with zero (e.g., "### ###"), one (e.g., "really ###") or two modifiers (e.g., "really not"). Modifiers were either the intensifier "really" or the negation "not". A sequence of dashes was used to indicate the absence of a modifier, e.g., "really ### good". Each of the eight adjectives was preceded by each of the nine possible combinations of modifiers: "### ###", "#### really", "really ###", "### not", "not ###", "really not", "not really", "really really" and "not not" ("not not" was included to achieve a full experimental design, even if it is not a frequent combination in natural language). To avoid possible differences in neural representation of phrases with and without syntactic/semantic composition, the condition with no modifiers ("### ###") was exclusively employed as a baseline comparison in the time-frequency analysis and was excluded from all other analyses. Each dimension (e.g., quality) was presented in two blocks, one block for each yes/no key orientation (8 blocks in total, see Procedure). Each phrase (e.g., "really really bad") was repeated four times within one block. Thus, the overall experiment comprised 576 trials. The order of phrases was randomized within each block for each participant. The order of blocks was randomized across participants within the first and second half of the experiment. The yes/no order was randomized across participants.

Procedure. Participants were familiarized with the linguistic stimuli through a short practice block that mimicked the structure of the experimental blocks. They were instructed to read affirmative or negated adjective phrases (e.g., "really really good", "really not bad") and derive the overall meaning of each adjective phrase, on a scale from 0 to 8, e.g., from "really really bad" to "really really good". Each trial started with a fixation cross (duration: 750 ms), followed by each phrase presented one word at a time, each word for 100 ms (inter-word-interval: 250 ms, **Fig.1B**). After each phrase, a fixation cross was presented for 1500 ms. A number (i.e., probe) was then presented, which did or did not correspond to the overall meaning of the adjective phrase on the scale. Participants were required to indicate whether the probe number correctly represented the meaning of the phrase on the scale (*yes/no* answer). The yes/no order was swapped halfway through the experiment. Responses had no time limit. If correct (+/- one step on the scale), a green fixation cross was presented; if incorrect, a red fixation cross was presented, and feedback was provided.
While performing the experiment, participants lay supine in a magnetically shielded room while continuous MEG data were recorded through a 157-channel whole-head axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan). Sampling rate was 1000 Hz, and online high-pass filter of 1 Hz and low-pass filter of 200 Hz were applied. Five electromagnetic coils were attached to the forehead of the participants and their position was measured twice, before the first and after the last block. Instructions, visual stimuli and visual feedback were back-projected onto a Plexiglas screen using a Hitachi projector. Stimuli were presented using Psychtoolbox v3 ([88]; www.psychtoolbox.org), running under MATLAB R2019a (MathWorks) on an Apple iMac model 10.12.6. Participants responded to the yes/no question with their index finger of their left and right hand, using a keypad. For each trial, we also collected accuracy and RTs. The overall duration of the MEG experiment was approximately 60 minutes.

Data analysis

Experiment 1 (and replication): RTs and mouse trajectories data.

The RTs and mouse trajectory analyses were limited to correct trials (group mean accuracy: 82%, SD: 13%), and RTs were limited within the range of participant median RTs ± 2 SD.
To evaluate differences in RTs between antonyms ("small", "cold", "ugly", "bad", "sad" vs. "big", "hot", "beautiful", "good", "happy", "fast", i.e., *low* vs. *high* poles in each scalar dimension), and between negated and affirmative phrases (e.g., "really really good" vs. "really not good"), and their interactions, median RTs of each participant were entered into 2 (*antonym*: low vs. high) x 2 (*negation*: negated vs. affirmative) repeated-measures ANOVA.
To evaluate differences in the final interpretations between antonyms in each scale, between negated and affirmative phrases, and their interactions, mean and standard deviation of the final responses of each participant were entered into a 2 (*antonym*: low vs. high) x 2 (*negation*: negated vs. affirmative) repeated-measures ANOVA. Post-hoc tests were conducted for significant interactions (correction = Holm). Effect sizes were calculated using partial eta squared (η_p^2).

To compare mouse trajectories over time across participants, we resampled participants' mouse trajectories at 100 Hz using linear interpolation, up to 2 seconds, to obtain 200 time points for each trial. Furthermore, trajectories were normalized between -1 and 1. For visualization purposes, we computed the median of trajectories across trials for each participant, dimension (e.g., quality), antonym (e.g., "bad") and modifier (e.g., "really not"), and at each timepoint.

Finally, to quantitatively evaluate how the interpretation of each phrase changed over time, for every participant we carried out regression analyses per each time point, for affirmative and negated phrases separately (for a similar approach, see [45]). The dependent variable was the mouse coordinate along the scale (note that the scale which was swapped in half of the trials was swapped back for data analysis purposes), and the predictor was whether the adjective was a low or high antonym (e.g., "bad" vs. "good"). To identify the time windows where predictors were significantly different from 0 at the group level, we performed permutation cluster tests on beta values (10,000 permutations) in the time window from the onset of the adjective up to 1.4 s from adjective onset (i.e., 2 s from the onset of word 1).

Experiment 2: Accuracy and RTs data.

To evaluate differences in accuracy between *low* and *high* antonyms ("bad", "cool", "quiet", "dark" vs. "good", "warm", "loud", "bright"), and between negated and affirmative phrases (e.g., "really really good" vs. "really not good"), and their interactions, mean accuracies in the yes/no task of each participant were entered into 2 (*antonym*: low vs. high) x 2 (*negation*: negated vs. affirmative) repeated-measures ANOVA.

The response time analysis was limited to correct trials. RTs outside the range of participant median RTs ± 2 SD were removed. To evaluate differences in RTs between *low* and *high* antonyms in each scale and between negated and affirmative phrases, and their interactions, median RTs of each participant in the yes/no task were entered into a 2 (*antonym*: low vs. high) x 2 (*negation*: negated vs. affirmative) repeated-measures ANOVA.

Experiment 2: MEG data.

Preprocessing.

MEG data preprocessing was performed using MNE-python [89] and Eelbrain (10.5281/zenodo.438193). First, bad channels (i.e., below the 3rd or above the 97th percentile across all channels, for more than 20% of the entire recording) were interpolated. The MEG responses were denoised by applying least square projections of the reference channels and removing the corresponding components from the data [90]. Denoised data were lowpass-filtered at 20 Hz for the decoding analyses and at 40 Hz for the time-frequency analyses. FastICA was used to decompose the signal into independent components, to visually inspect and remove artifacts related to eye-blinks, heartbeat and external noise sources. MEG recordings were then epoched into epochs of -300 ms and 2550 ms around the onset of the first, second, or third word (or probe) for the decoding analyses, and into epochs of -800 and 3000 ms around the onset of the first word for the time-frequency analyses (and then cut between -300 and 2550 ms for group analyses). Note that, for visualization purposes, only 1700 ms from the onset of the first word (i.e., 1000 ms from adjective onset) were included in most figures (as no significant results were observed for control analyses run for later time windows). Finally, epochs with amplitudes greater than an absolute threshold of 3000 fT were removed and a baseline between -300 to 0 ms was applied to all epochs.

Source reconstruction.

Structural magnetic resonance images (MRIs) were collected for 10 out of 26 participants. For the remaining 16 participants, we manually scaled and co-registered the “fsaverage” brain to the participant’s head-digitalized shape and fiducials [89,91].

For every participant, an ico-4 source space was computed, containing 2562 vertices per hemisphere and the forward solution was calculated using the Boundary Element Model (BEM). A noise covariance matrix was estimated from the 300 ms before the onset of the first word. The inverse operator was created and applied to the neuromagnetic data to estimate the source time courses at each vertex using dynamic statistical parametric mapping (dSPM: [92]). The results were then morphed to the ico-5 “fsaverage” brain, yielding to time courses for 10242 vertices per hemisphere.

We then estimated the magnitude of the activity at each vertex (signal to noise ratio: 3, lambda2: 0.11, with orientation perpendicular to the cortical surface), which was used in the decoding analyses (*Spatial decoders*).

Decoding analyses.

Decoding analyses were limited to correct trials and were performed with the MNE [89] and Scikit-Learn packages [47]. First, X (or the selected principal components) were set to have zero mean and unit variance (i.e., using a standard scaler). Second, we fitted a l2 linear estimator to a subset of the epochs (training set, X_{train}) and estimated y on a separate group of epochs (test set, $\hat{y}_{test}$). We then computed the accuracy (AUC, see below) of the decoder, by comparing $\hat{y}_{test}$ with the ground truth y. For this analysis, we used the default values provided by the Scikit-Learn package and set the class-weight parameter to "balanced".

Temporal decoders. Temporal decoding analyses were performed in sensor-space. Before fitting the estimators, linear dimensionality reduction (principal component analysis, PCA) was performed on the channel amplitudes to project them to a lower dimensional space (i.e., to new virtual channels that explained more than 99% of the feature variance). We then fitted the linear estimator on each participant separately, across all selected components, at each time-point separately. Time was subsampled to 100 Hz. We then employed a 5-fold stratified cross-validation (or 10-fold, depending on the number of trials per class), that fitted the linear estimator to 80% (or 90%) of the epochs and generated predictions on 20% (or 10%) of the epochs, while keeping the distributions of the training and test set maximally homogeneous. This decoding approach was used for analyses of **Fig.3B, Fig.3C, Fig.4A, Fig.4C** and decoding approach (ii) in **Fig.5C**. To investigate whether the representation of antonyms was comparable between affirmative and negated phrases, in a different set of analyses (i.e., decoding approach (i), **Fig.5B**) we fitted the linear estimator to all epochs corresponding to affirmative phrases and generated predictions on all epochs corresponding to negated phrases. In both decoding approaches, accuracy and probability estimates for each class were then computed. Decoding accuracy is summarized with an empirical area under the curve (rocAUC, 0 to 1, chance at 0.5).

At the group level, we extracted the clusters of time where AUC across participants was significantly higher than chance using a one-sample permutation cluster test, as implemented in MNE-python (10000 permutations [93]). We performed separate permutation cluster tests for the following time windows: -700 to -350 ms from adjective onset (i.e., word 1), -350 to 0 ms from adjective onset (i.e., word 2), 0 to 500 ms from adjective onset (i.e., time window for lexical-semantic processes [65,66]) and 500 to 1000 ms from adjective onset (i.e., to account for potential later processes).

Expected outcome for the effect of negation on the representation of antonyms. Temporal decoding approach (i) and (ii) described above allow us to make specific predictions about the effect of negation on the representation of antonyms (**Fig.5A**).

Approach (i) train set: affirmative phrases; test set: negated phrases. For our results to support predictions (1) *No effect* or (2) *Mitigation*, this decoding approach should show probability estimates of high and low adjectives significantly above the computed chance level and in the direction of the respective classes, indicating that the initial representation of adjectives in negated phrases is similar to that in affirmative phrases (left column, first and second row under *decoding approach* in **Fig.5A**). Conversely, for our results to support prediction (3) *Inversion*, this decoding approach should show probability estimates of high and low adjectives significantly above the computed chance level but in the direction of the opposite classes (i.e., swapped), as adjective representations would be systematically inverted in negated phrases (left column, third row under *decoding approach* in **Fig.5A**). Finally, we should observe at chance probability estimates in the case of (4) *Change*, where adjective representations in negated phrases are not predictable from the corresponding representations in affirmative phrases (left column, fourth row under *decoding approach* in **Fig.5A**).

Approach (ii) train set: affirmative and negated phrases together; test set: affirmative and negated phrases separately. This decoding analysis allows us to disentangle predictions (1) *No effect* from (2) *Mitigation*. For the results of this analysis to support prediction (1) *No effect*, we should observe quantitatively comparable probability estimates in affirmative and negated phrases, suggesting that negation does not change the representation of adjectives (right column, first row under *decoding approach* in **Fig.5A**). Conversely, in support of prediction (2) *Mitigation*, we should observe significantly reduced probability estimates for negated relative to affirmative phrases, suggesting less robust differences between low and high antonyms in negated phrases (right column, second row under *decoding approach* in **Fig.5A**). The outcome of predictions (3) *Inversion* and (4) *Change* would be at chance probability estimates (as the model is trained on different representations within the same class; right column, third and fourth row under *decoding approach* in **Fig.5A**).

Spatial decoders. Spatial decoding analyses were performed in source-space. We fitted each estimator on each participant separately, across 50 to 650 ms time samples relative to the onset of the adjective (to include the three significant time windows that emerge from the temporal decoding analysis in **Fig.3B**), at each brain source separately, after morphing individual participant's source estimates to the ico-5 "fsaverage" common reference space. We employed a 5-fold stratified cross-validation, which fitted the linear estimator to 80% of the epochs and generated predictions on 20% of the epochs, while keeping the distributions of the training and test set maximally homogeneous. Decoding accuracy is summarized with an empirical area under the curve (AUC, 0 to 1, chance at 0.5). At the group level, we extracted the brain areas where the AUC across participants was significantly higher than chance, using a one-sample permutation cluster test as implemented in MNE-python (10000 permutations; adjacency computed from the "fsaverage" brain [93]).

Time-frequency analysis.

We extracted time-frequency power of the epochs (-800 to 3000 ms from the onset of word 1) using Morlet wavelets of 3 cycles per frequency, in frequencies between 3.9 and 37.2 Hz, logarithmically spaced (19 frequencies overall). Power estimates where then cut between -300 and 2550 ms from onset of word 1 and baseline corrected using a window of -300 to -100 ms from the onset of word 1, by subtracting the mean of baseline values and dividing by the mean of baseline values (mode = 'percent'). Power in the low-beta frequency range (12 to 20 Hz) and in the high-beta frequency range (21 to 30 Hz [57,79]) was averaged to obtain a time course of power in low and high-beta rhythms. We then subtracted the beta power of affirmative phrases from that of negated phrases. At the group level, we extracted the clusters of time where this difference in power across participants was significantly greater than 0, using a one-sample permutation cluster test as implemented in MNE-python (10000 permutations [93]). We performed separate permutation cluster tests in the same time windows used for the decoding analysis: -700 to -350 ms, -350 to 0 ms, 0 to 500 ms, and 500 to 1000 ms from the onset of the adjective (note that no significant differences were observed in analyses ran for time windows after 1000 ms). We then computed the induced power in source space (method: dSPM and morphing individual participant's source estimates to the ico-5 "fsaverage" reference space) for the significant clusters of time in the low- and high-beta range separately and averaged over time. At the group level, we extracted the brain areas where the power difference across participants was significantly greater than 0, using a one-sample permutation cluster test as implemented in MNE-python (10000 permutations; adjacency computed from the "fsaverage" brain [93]).

References

1. Ding, N., Melloni, L., Zhang, H., Tian, X. & Poeppel, D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience* **19**, 158–164 (2015).
2. Fedorenko, E. *et al.* Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E6256–E6262 (2016).
3. Martin, A. E. & Baggio, G. Modelling meaning composition from formalism to mechanism. 1–7 (2019).
4. Matchin, W. & Hickok, G. The Cortical Organization of Syntax. *Cerebral Cortex* **30**, 1481–1498 (2020).
5. Oseki, Y. & Marantz, A. Modeling morphological processing in human magnetoencephalography. *Proceedings of the Society for Computation in Linguistics* **3**, (2020).
6. Pallier, C., Devauchelle, A.-D. & Dehaene, S. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences* **108**, 2522–2527 (2011).
7. Pylkkänen, L. The neural basis of combinatory syntax and semantics. *Science* **366**, 62–66 (2019).
8. Ziegler, J. & Pylkkänen, L. Scalar adjectives and the temporal unfolding of semantic composition: An MEG investigation. *Neuropsychologia* **89**, 161–171 (2016).
9. Tian, Y., Ferguson, H. & Breheny, R. Processing negation without context – why and when we represent the positive argument. *Language, Cognition and Neuroscience* **31**, 683–698 (2016).
10. Tian, Y., Breheny, R. & Ferguson, H. J. Why we simulate negated information: A dynamic pragmatic account. *Quarterly Journal of Experimental Psychology* **63**, 2305–2312 (2010).
11. Giora, R. Anything negatives can do affirmatives can do just as well, except for some metaphors. *Journal of Pragmatics* **38**, 981–1014 (2006).
12. Horn, L. R. *A natural history of negation*. (University of Chicago Press, 1989).
13. Ettinger, A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics* **8**, 34–48 (2020).
14. Dale, R. & Duran, N. D. The cognitive dynamics of negated sentence verification. *Cognitive Science* **35**, 983–996 (2011).
15. Darley, E. J., Kent, C. & Kazanina, N. A 'no' with a trace of 'yes': A mouse-tracking study of negative sentence processing. *Cognition* **198**, 104084 (2020).
16. Dudschig, C. & Kaup, B. How does 'not left' become 'right'? Electrophysiological evidence for a dynamic conflict-bound negation processing account. *Journal of Experimental Psychology: Human Perception and Performance* **44**, 716–728 (2018).

17. Just, M. A. & Carpenter, P. A. Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior* **10**, 244–253 (1971).
18. Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A. & Ldtke, J. Experiential simulations of negated text information. *Quarterly Journal of Experimental Psychology* **60**, 976–990 (2007).
19. Dudschig, C., Kaup, B., Liu, M. & Schwab, J. The processing of negation and polarity: An overview. *Journal of Psycholinguistic Research* **50**, 1199–1213 (2021).
20. Sherman, M. A. Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior* **15**, 143–157 (1976).
21. Kaup, B. Negation and its impact on the accessibility of text information. *Memory and Cognition* **29**, 960–967 (2001).
22. Kaup, B. & Zwaan, R. A. Effects of negation and situational presence on the accessibility of text information. *Journal of Experimental Psychology: Learning Memory and Cognition* **29**, 439–446 (2003).
23. MacDonald, M. C. & Just, M. A. Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**, 633–642 (1989).
24. Carpenter, P. A. & Just, M. A. Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review* **82**, 45–73 (1975).
25. Clark, H. H. & Chase, W. G. On the process of comparing sentences against pictures. *Cognitive Psychology* **3**, 472–517 (1972).
26. Lüdtke, J., Friedrich, C. K., De Filippis, M. & Kaup, B. Event-related potential correlates of negation in a sentence-picture verification paradigm. *Journal of Cognitive Neuroscience* **20**, 1355–1370 (2008).
27. Kaup, B. & Dudschig, C. Understanding negation: Issues in the processing of negation. in *The Oxford Handbook of Negation* (eds. Déprez, V. & Espinal, M. T.) 634–655 (Oxford University Press, 2020).
28. Papeo, L. & de Vega, M. The neurobiology of lexical and sentential negation. *The Oxford Handbook of Negation* 739–756 (2020).
29. Papeo, L., Hochmann, J.-R. & Battelli, L. The default computation of negated meanings. *Journal of Cognitive Neuroscience* **28**, 1980–1986 (2016).
30. Lyons, J. *Linguistic semantics: An introduction*. (Cambridge University Press, 1995).
31. Mayo, R., Schul, Y. & Burnstein, E. 'I am not guilty' vs 'I am innocent': Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology* **40**, 433–449 (2004).
32. Orenes, I., Beltrán, D. & Santamaría, C. How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language* **74**, 36–45 (2014).

33. van Gaal, S. *et al.* Can the meaning of multiple words be integrated unconsciously? *Phil. Trans. R. Soc. B* **369**, 20130212 (2014).
34. Bartoli, E. *et al.* The disembodiment effect of negation: Negating action-related sentences attenuates their interference on congruent upper limb movements. *Journal of Neurophysiology* **109**, 1782–1792 (2013).
35. Beltrán, D., Morera, Y., García-Marco, E. & De Vega, M. Brain inhibitory mechanisms are involved in the processing of sentential negation, regardless of its content. Evidence from EEG theta and beta rhythms. *Frontiers in Psychology* **10**, 1–14 (2019).
36. Beltrán, D., Liu, B. & de Vega, M. Inhibitory mechanisms in the processing of negations: A neural reuse hypothesis. *Journal of Psycholinguistic Research* **50**, 1243–1260 (2021).
37. De Vega, M. *et al.* Sentential negation might share neurophysiological mechanisms with action inhibition. Evidence from frontal theta rhythm. *Journal of Neuroscience* **36**, 6002–6010 (2016).
38. Djokic, V., Maillard, J., Bulat, L. & Shutova, E. Modeling affirmative and negated action processing in the brain with lexical and compositional semantic models. 5155–5165 (2019).
39. Gallese, V. & Lakoff, G. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology* **22**, 455–479 (2005).
40. Tettamanti, M. *et al.* Negation in the brain: Modulating action representations. *NeuroImage* **43**, 358–367 (2008).
41. Tomasino, B., Weiss, P. H. & Fink, G. R. To move or not to move: Imperatives modulate action-related verb processing in the motor system. *Neuroscience* **169**, 246–258 (2010).
42. Bianchi, I., Savardi, U., Burro, R. & Torquati, S. Negation and psychological dimensions. *Journal of Cognitive Psychology* **23**, 275–301 (2011).
43. Colston, H. L. 'Not good' is 'bad,' but 'not bad' is not 'good': an analysis of three accounts of negation asymmetry. *Discourse Processes* **28**, 237–256 (1999).
44. Fraenkel, T. & Schul, Y. The meaning of negated adjectives. *Intercultural Pragmatics* **5**, 517–540 (2008).
45. Dotan, D. & Dehaene, S. How do we convert a number into a finger trajectory? *Cognition* **129**, 512–529 (2013).
46. Maldonado, M., Dunbar, E. & Chemla, E. Mouse tracking as a window into decision making. *Behav Res* **51**, 1085–1101 (2019).
47. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
48. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun Biol* **5**, 134 (2022).

49. Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex* **19**, 2767–2796 (2009).
50. Caucheteux, C., Gramfort, A. & King, J.-R. Disentangling syntax and semantics in the brain with deep networks. (2021).
51. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
52. Lau, E. F., Gramfort, A., Hämäläinen, M. S. & Kuperberg, G. R. Automatic semantic facilitation in anterior temporal cortex revealed through multimodal neuroimaging. *Journal of Neuroscience* **33**, 17174–17181 (2013).
53. Lambon Ralph, M. A., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience* **18**, 42–55 (2016).
54. Hagoort, P., Hald, L., Bastiaansen, M. & Petersson, K. M. Integration of word meaning and world knowledge in language comprehension. *Science* **304**, 438–441 (2004).
55. Popham, S. F. *et al.* Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat Neurosci* **24**, 1628–1636 (2021).
56. Parrish, A. & Pylkkänen, L. Conceptual combination in the LATL with and without syntactic composition. *Neurobiology of Language* **3**, 46–66 (2022).
57. Weiss, S. & Mueller, H. M. 'Too many betas do not spoil the broth': The role of beta brain oscillations in language processing. *Frontiers in Psychology* **3**, 1–15 (2012).
58. Wagner, J., Wessel, J. R., Ghahremani, A. & Aron, A. R. Establishing a right frontal beta signature for stopping action in scalp EEG: Implications for testing inhibitory control in other task contexts. *Journal of Cognitive Neuroscience* **30**, 107–118 (2018).
59. Alemanno, F. *et al.* Action-related semantic content and negation polarity modulate motor areas during sentence reading: An event-related desynchronization study. *Brain Research* **1484**, 39–49 (2012).
60. Aravena, P. *et al.* Grip force reveals the context sensitivity of language-induced motor activity during "action words" processing: evidence from sentential negation. *PLoS ONE* **7**, e50287 (2012).
61. Foroni, F. & Semin, G. R. Comprehension of action negation involves inhibitory simulation. *Frontiers in Human Neuroscience* **7**, 1–7 (2013).
62. Ghio, M., Haegert, K., Vaghi, M. M. & Tettamanti, M. Sentential negation of abstract and concrete conceptual categories: A brain decoding multivariate pattern analysis study. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**, 7–10 (2018).

63. Liuzza, M. T., Candidi, M. & Aglioti, S. M. Do not resonate with actions: Sentence polarity modulates cortico-spinal excitability during action-related sentence reading. *PLoS ONE* **6**, 38–41 (2011).
64. Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F. & Marslen-Wilson, W. D. The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage* **30**, 1383–1400 (2006).
65. Kutas, M. & Federmeier, K. D. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology* **62**, 621–647 (2011).
66. Pulvermüller, F., Shtyrov, Y. & Hauk, O. Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language* **110**, 81–94 (2009).
67. Pulvermüller, F., Assadollahi, R. & Elbert, T. Neuromagnetic evidence for early semantic access in word recognition. *European Journal of Neuroscience* **13**, 201–205 (2001).
68. Teige, C. *et al.* Dynamic semantic cognition: Characterising coherent and controlled conceptual retrieval through time using magnetoencephalography and chronometric transcranial magnetic stimulation. *Cortex* **103**, 329–349 (2018).
69. Zhang (张琳敏), L. & Pylkkänen, L. Semantic composition of sentences word by word: MEG evidence for shared processing of conceptual and logical elements. *Neuropsychologia* **119**, 392–404 (2018).
70. Fyshe, A., Sudre, G., Wehbe, L., Rafidi, N. & Mitchell, T. M. The lexical semantics of adjective–noun phrases in the human brain. *Human Brain Mapping* **40**, 4457–4469 (2019).
71. Nieuwland, M. S. & Kuperberg, G. R. When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science* **19**, 1213–1218 (2008).
72. Palaz, B., Rhodes, R. & Hestvik, A. Informative use of "not" is N400-blind. *Psychophysiology* **57**, (2020).
73. Xiang, M., Grove, J. & Giannakidou, A. Semantic and pragmatic processes in the comprehension of negation: An event related potential study of negative polarity sensitivity. *Journal of Neurolinguistics* **38**, 71–88 (2016).
74. Grice, H. P. Logic and Conversation. in *Syntax and Semantics* vol. 3 41–58 (New York: Academic Press, 1975).
75. Bastiaansen, M. C. M., van der Linden, M., ter Keurs, M., Dijkstra, T. & Hagoort, P. Theta responses are involved in lexical—semantic retrieval during language processing. *Journal of Cognitive Neuroscience* **17**, 530–541 (2005).
76. Luo, Y., Zhang, Y., Feng, X. & Zhou, X. Electroencephalogram oscillations differentiate semantic and prosodic processes during sentence reading. *Neuroscience* **169**, 654–664 (2010).
77. Supp, G. G. *et al.* Lexical memory search during N400: cortical couplings in auditory comprehension: *NeuroReport* **15**, 1209–1213 (2004).

78. Weiss, S. & Rappelsberger, P. EEG coherence within the 13–18 Hz band as a correlate of a distinct lexical organisation of concrete and abstract nouns in humans. *Neuroscience Letters* **209**, 17–20 (1996).

79. Schaller, F., Weiss, S. & Müller, H. M. EEG beta-power changes reflect motor involvement in abstract action language processing. *Brain and Language* **168**, 95–105 (2017).

80. Pinheiro-Chagas, P., Dotan, D., Piazza, M. & Dehaene, S. Finger tracking reveals the covert stages of mental arithmetic. *Open Mind* **1**, 30–41 (2017).

81. Peer, E., Brandimarte, L., Samat, S. & Acquisti, A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* **70**, 153–163 (2017).

82. Simcox, T. & Fiez, J. A. Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods* **46**, 95–111 (2014).

83. Gwilliams, L. & King, J. R. Recurrent processes support a cascade of hierarchical decisions. *eLife* **9**, 1–20 (2020).

84. King, J. R., Pescetelli, N. & Dehaene, S. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron* **92**, 1122–1134 (2016).

85. Balota, D. A. *et al.* The English Lexicon Project. *Behavior Research Methods* **39**, 445–459 (2007).

86. Schiller, N. O. *et al.* Solving the problem of double negation is not impossible: electrophysiological evidence for the cohesive function of sentential negation. *Language, Cognition and Neuroscience* **32**, 147–157 (2017).

87. Chen, D. L., Schonger, M. & Wickens, C. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* **9**, 88–97 (2016).

88. Brainard, D. H. The Psychophysics Toolbox. *Spatial Vision* **10**, 433–436 (1997).

89. Gramfort, A. *et al.* MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience* **7**, 1–13 (2013).

90. Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y. & Ochiai, M. Reduction of non-periodic environmental magnetic noise in MEG measurement by Continuously Adjusted Least squares Method. *IEEE Transactions on Applied Superconductivity* **11**, 669–672 (2001).

91. Andersen, L. M. Group analysis in MNE-python of evoked responses from a tactile stimulation paradigm: A pipeline for reproducibility at every step of processing, going from individual sensor space representations to an across-group source space representation. *Frontiers in Neuroscience* **12**, (2018).

92. Dale, A. M. *et al.* Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* **26**, 55–67 (2000).

93. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* **164**, 177–190 (2007).

Acknowledgements

This work was supported by the Leon Levy Foundation (A.Z.), the European Union's Horizon 2020 research and innovation program under grant agreement No 660086 (J.R.K.), the Bettencourt-Schueller Foundation (J.R.K.), the Fondation Roger de Spoelberch (J.K.R.), the Philippe Foundation (J.R.K.), the FrontCog grant ANR-17-EURE-0017 (J.R.K.), and the Ernst Struengmann Foundation (D.P.).

Author contributions

AZ, PR, JRK, and DP conceptualized the experiment; AZ, PR, and WML collected the data; AZ analyzed the data; PR, LG, and JRK contributed to analysis; AZ wrote the paper; AZ, PR, LG, JRK, and DP discussed the results and edited the paper.

Competing interests

The authors declare no competing interests.

Data and materials availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the corresponding author.

Supplementary Materials

Tables

List of linguistic stimuli employed in Experiment 1 (behavior)					
### ###	small	really really	small	not not	small
### ###	big	really really	big	not not	big
### ###	cold	really really	cold	not not	cold
### ###	hot	really really	hot	not not	hot
### ###	ugly	really really	ugly	not not	ugly
### ###	beautiful	really really	beautiful	not not	beautiful
### ###	bad	really really	bad	not not	bad
### ###	good	really really	good	not not	good
### ###	sad	really really	sad	not not	sad
### ###	happy	really really	happy	not not	happy
### ###	slow	really really	slow	not not	slow
### ###	fast	really really	fast	not not	fast
### really	small	### not	small	really not	small
### really	big	### not	big	really not	big
### really	cold	### not	cold	really not	cold
### really	hot	### not	hot	really not	hot
### really	ugly	### not	ugly	really not	ugly
### really	beautiful	### not	beautiful	really not	beautiful
### really	bad	### not	bad	really not	bad
### really	good	### not	good	really not	good
### really	sad	### not	sad	really not	sad
### really	happy	### not	happy	really not	happy
### really	slow	### not	slow	really not	slow
### really	fast	### not	fast	really not	fast
really ###	small	not ###	small	not really	small
really ###	big	not ###	big	not really	big
really ###	cold	not ###	cold	not really	cold
really ###	hot	not ###	hot	not really	hot
really ###	ugly	not ###	ugly	not really	ugly
really ###	beautiful	not ###	beautiful	not really	beautiful
really ###	bad	not ###	bad	not really	bad
really ###	good	not ###	good	not really	good
really ###	sad	not ###	sad	not really	sad
really ###	happy	not ###	happy	not really	happy
really ###	slow	not ###	slow	not really	slow
really ###	fast	not ###	fast	not really	fast

Table S1. Comprehensive list of the 108 stimuli used in the behavioral experiment, color coded for each experimental condition; purple: low adjectives, orange: high adjectives; green: affirmative phrases, red: negated phrases.

List of linguistic stimuli employed in Experiment 2 (MEG)					
### ###	quiet	really really	quiet	not not	quiet
### ###	loud	really really	loud	not not	loud
### ###	cool	really really	cool	not not	cool
### ###	warm	really really	warm	not not	warm
### ###	dark	really really	dark	not not	dark
### ###	bright	really really	bright	not not	bright
### ###	bad	really really	bad	not not	bad
### ###	good	really really	good	not not	good
### really	quiet	### not	quiet	really not	quiet
### really	loud	### not	loud	really not	loud
### really	cool	### not	cool	really not	cool
### really	warm	### not	warm	really not	warm
### really	dark	### not	dark	really not	dark
### really	bright	### not	bright	really not	bright
### really	bad	### not	bad	really not	bad
### really	good	### not	good	really not	good
really ###	quiet	not ###	quiet	not really	quiet
really ###	loud	not ###	loud	not really	loud
really ###	cool	not ###	cool	not really	cool
really ###	warm	not ###	warm	not really	warm
really ###	dark	not ###	dark	not really	dark
really ###	bright	not ###	bright	not really	bright
really ###	bad	not ###	bad	not really	bad
really ###	good	not ###	good	not really	good

Table S2. Comprehensive list of the 72 stimuli used in the MEG experiment, color coded for each experimental condition; purple: low adjectives, orange: high adjectives; green: affirmative phrases, red: negated phrases. Note that the condition with no modifiers ("### ###") was only employed as a baseline condition in the time-frequency analysis.

Figures

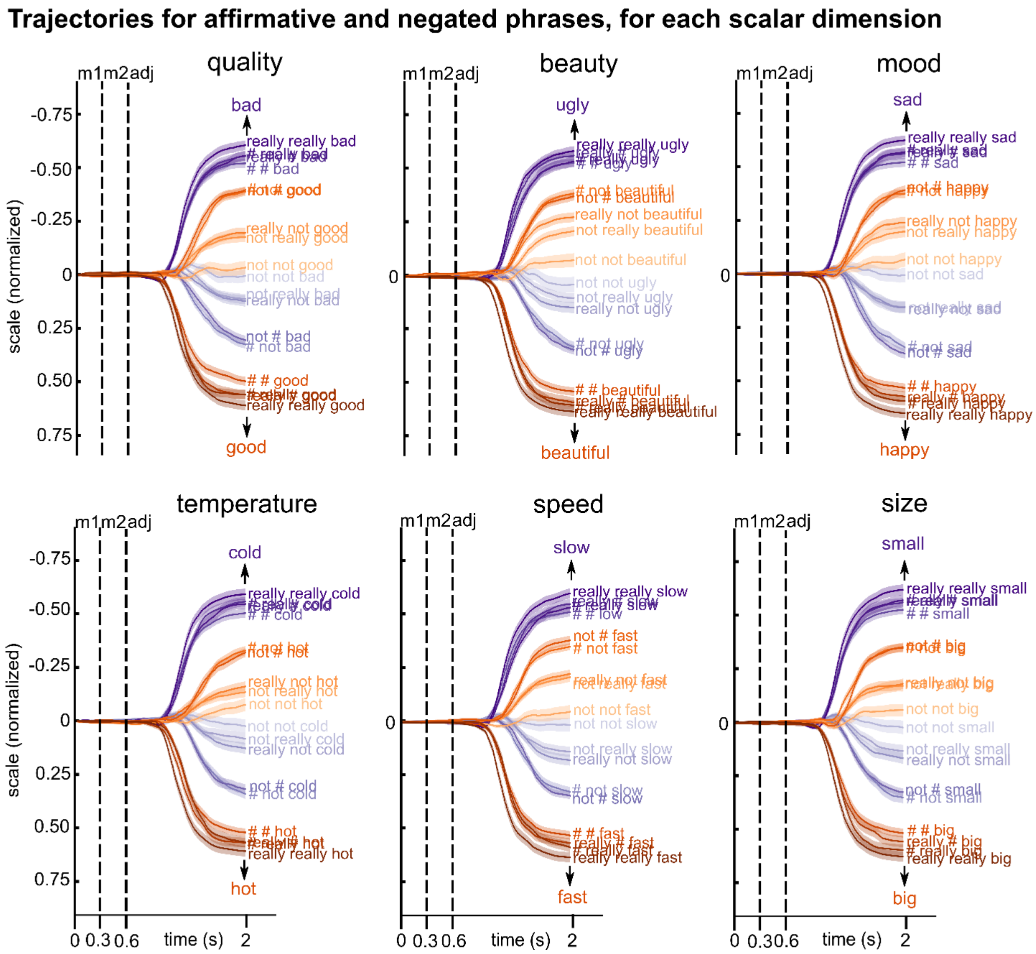


Fig. S1. Trajectories for each scalar dimension.

Behavioral trajectories for low (purples) and high (oranges) antonyms over time, for each scalar dimension (i.e., quality, beauty, mood, temperature, speed and size), for each modifier (shades of orange and purple), and for affirmative and negated phrases. Black vertical dashed lines indicate the presentation onset of each word: modifier1, modifier2 and adjective.

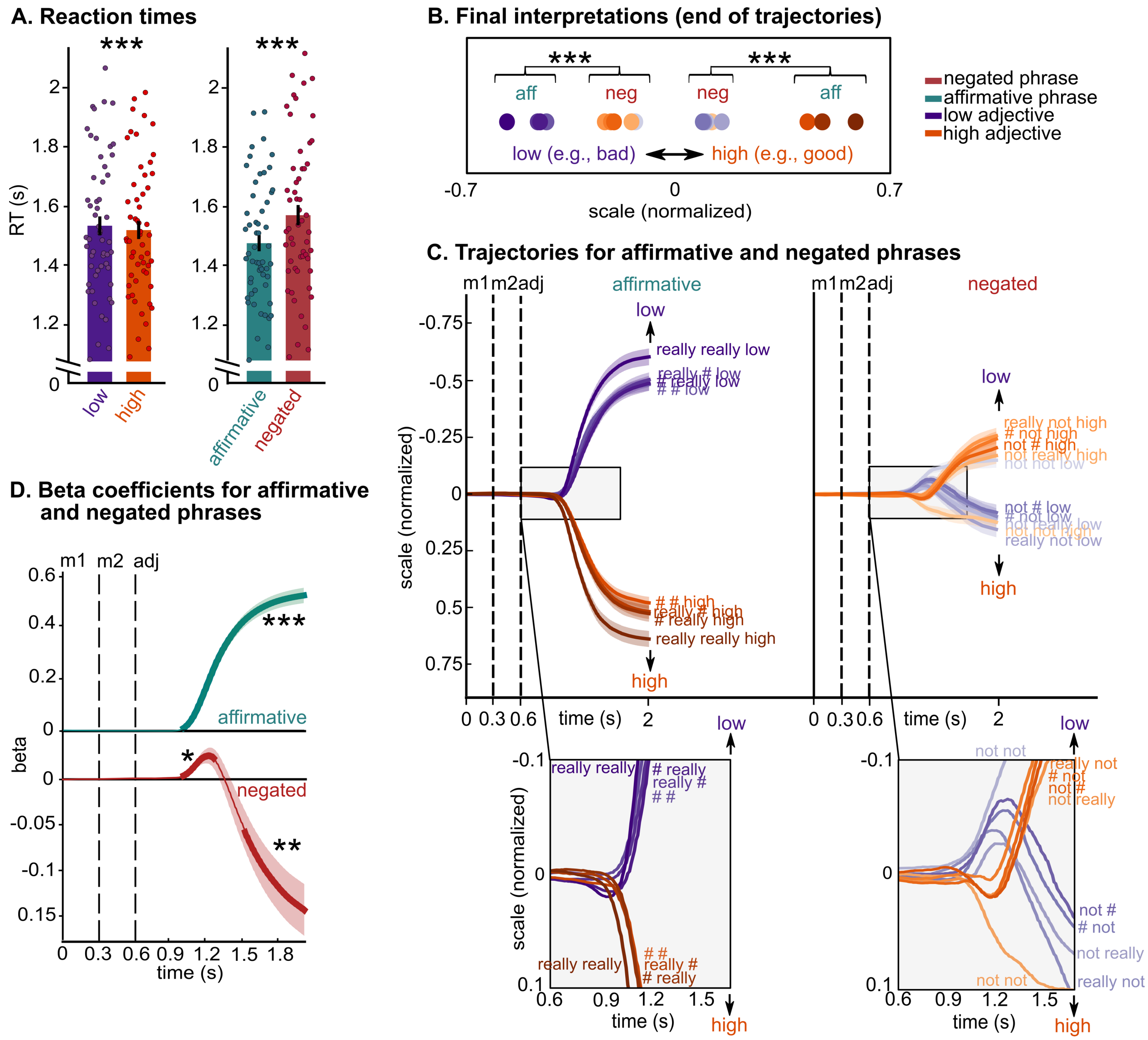


Fig. S2. Replication of Experiment 1, without feedback on interpretation.

A new group of 60 participants (37 females; mean age = 19.26 years; range 18-23 years) completed the online mouse tracking experiment. Participants were recruited via the platform SONA (a platform for students' recruitment). All participants were native English speakers with self-reported normal hearing and no neurological deficits. 59 participants were right-handed. Participants were granted university credits for taking part in the study, which was performed online. All participants provided written informed consent, as approved by the local institutional review board (New York University's Committee on Activities Involving Human Subjects). The data of 5 participants were excluded from the data analysis due to (i) number of "incorrect" feedback based on the warnings > 30%, (ii) mean RTs > 2SD from the group mean, or (iii) response trajectory always ending within 1/4 from the center of the scale, regardless of condition (i.e., participants who did not pay attention to the instructions of the task). Thus, 55 participants were included in the analysis. The experimental

procedure was the same as that of Experiment 1, except that no feedback was provided to participants based on the final interpretation, but only if the cursor's movement violated the warnings provided during the familiarization phase (e.g., "you crossed the vertical borders", see Procedure of Experiment 1). We performed the same data analyses performed for Experiment 1: (**A**) *Reaction times:* To evaluate the specific effect of antonyms and of negation, we performed a 2 (*antonym*: low vs high) x 2 (*negation*: negated vs affirmative) repeated-measures ANOVA. The results reveal a significant main effect of antonyms ($F(1,54) = 36.90$, $p < 0.001$, $\eta_p^2 = 0.40$) and a significant main effect of negation ($F(1,54) = 73.04$, $p < 0.001$, $\eta_p^2 = 0.57$). Moreover, a significant crossover interaction between antonyms and negation was found ($F(1,54) = 16.40$, $p < 0.001$, $\eta_p^2 = 0.23$). These results replicate Experiment 1, showing that participants were faster for high adjectives (e.g., "good") than for low adjectives (e.g., "bad") and for affirmative phrases (e.g., "really really good") than for negated phrases (e.g., "really not good"). A further analysis including the number of modifiers as factor (i.e., *complexity*) indicates that participants were faster for phrases with two modifiers, e.g., "not really", than phrases with one modifier, e.g., "not ###" ($F(1,54) = 28.87$, $p < 0.001$, $\eta_p^2 = 0.35$, especially in affirmative phrases: complexity by negation interaction $F(1,54) = 6.26$, $p = 0.015$, $\eta_p^2 = 0.10$), again replicating results of Experiment 1. Bars represent the participants' mean ± SEM and dots represent individual participants.
(**B**) *Continuous mouse trajectories:* To investigate how negation changes the interpretation of scalar adjectives, we performed a 2 (*antonym*: low vs high) x 2 (*negation*: negated vs affirmative) repeated-measures ANOVA for participants' final interpretations (filled circles, purple = low, orange = high, averaged across dimensions and participants), which revealed a significant main effect of antonyms ($F(1,54) = 166.40$, $p < 0.001$, $\eta_p^2 = 0.47$), a significant main effect of negation ($F(1,54) = 48.62$, $p < 0.001$, $\eta_p^2 = 0.47$), and a significant interaction between antonyms and negation ($F(1,54) = 210.13$, $p < 0.001$, $\eta_p^2 = 0.80$). Post-hoc tests show that the final interpretation of negated phrases was located at a more central portion of the semantic scale than that of affirmative phrases (affirmative low < negated high, and affirmative high > negated low, $p_{holm} < 0.001$), indicating that negation never inverts the interpretation of adjectives to that of their antonyms. Results also show that the final interpretations of negated phrases was significantly more variable (measured as standard deviations) than that of affirmative phrases ($F(1,54) = 15.43$, $p < 0.001$, $\eta_p^2 = 0.22$). These results replicate Experiment 1. (**C**) and (**D**). To quantify the degree of deviation towards each side of the scale, we performed regression analyses with antonyms as the predictor and mouse trajectories as the dependent variable. Trials with "not not" were not included in this analysis as, in this experiment, the trajectories pattern was different compared to the other conditions with negation. Our results indicate that, while mouse trajectories of affirmative phrases

branched towards either side of the scale and remained on that side until the final interpretation (lines in the left, gray, zoomed-in panel in **C**), the trajectories of negated phrases first deviated towards the side of the adjective and then towards the side of the antonym (lines in the right, gray, zoomed-in panel in **C**). The results of the regression analyses show that (1) in affirmative phrases, betas are positive (i.e., mouse trajectories moving towards the adjective) starting from 400 ms from the adjective onset ($p < 0.001$, green line in **D**); and that (2) in negated phrases, betas are positive (i.e., mouse trajectories moving towards the adjective) between 400 and 650 ms from the adjective onset ($p = 0.02$), and only became negative (i.e., mouse trajectories moving towards the antonym) from 910 ms from the adjective onset ($p = 0.003$, i.e., red line in **D**). Thicker lines indicate significant time windows. These results again replicate Experiment 1. For panels C and D: black vertical dashed lines indicate the presentation onset of each word: modifier 1, modifier 2 and adjective; each line and shading represent participants' mean ± SEM; *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

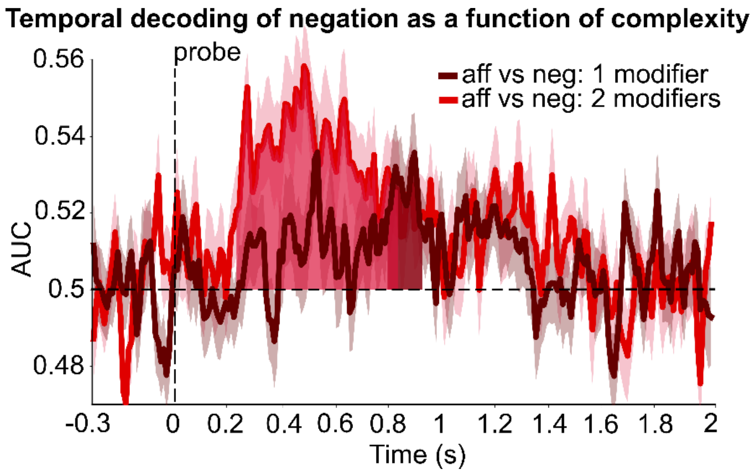


Fig. S3. Temporal decoding of negation as a function of number of modifiers (i.e., complexity), time-locked to the onset of the probe.

Decoding accuracy of negation over time, as a function of the number of modifiers (1 modifier: dark red line and shading; 2 modifiers: light red line and shading). Significant time windows are indicated by dark red (1 modifier) and light red (2 modifiers) areas. These results show that we could significantly decode the difference between affirmative and negated phrases between 230 and 930 ms after the onset of the probe, especially when the phrase included two modifiers (1 modifier: between 790 and 930 ms: $p < 0.001$; 2 modifiers: between 230 and 840 ms: $p < 0.001$). This suggests that the representation of modifiers is reactivated at the stage when participants have to perform the yes/no task. AUC = area under the receiver operating characteristic curve, chance = 0.5 (black dashed horizontal line); the black vertical dashed line indicates the presentation onset of the probe; aff = affirmative, neg = negated; each line and shading represent participants mean ± SEM.

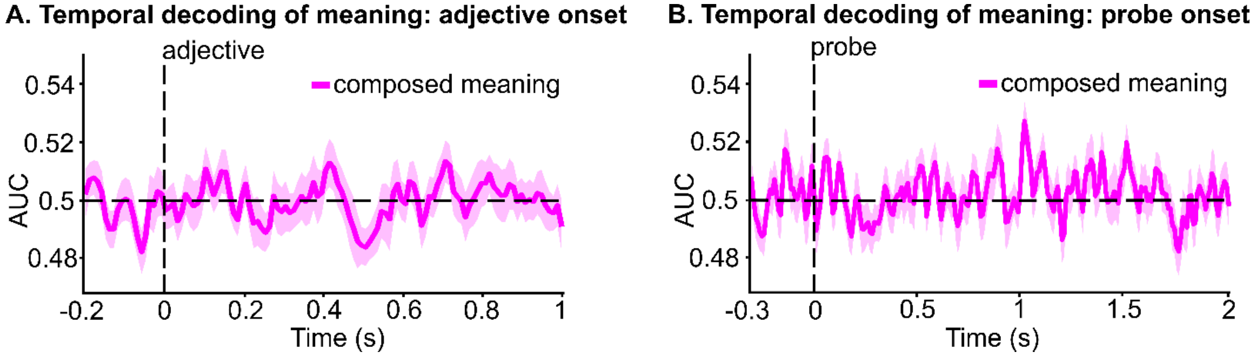


Fig. S4. Temporal decoding of composed meaning.

We trained estimators on phrases where the predicted composed meaning was “low” vs. “high” in 90% of the trials and computed the accuracy of the model in predicting the representation of the meaning “low” vs. “high” in the remaining 10% of the trials. For instance, for the *quality* dimension, classes are: [0: *bad*] “### really bad”, “really ### bad”, “really really bad”, “### not good”, “not ### good”, “not not good”, “really not good”, “not really good”; and [1: *good*] “### really good”, “really ### good”, “really really good”, “### not bad”, “not ### bad”, “not not bad”, “really not bad”, “not really bad”. The composed meaning was derived from the behavioral results of Experiment 1. (**A**) Temporal decoding analyses time-locked to the onset of the adjective do not reveal any significant temporal cluster, suggesting that negation does not invert the representation of the adjective to that of its antonym (e.g., “bad” to “good”), as would be predicted by prediction (3) *Inversion*. (**B**) Temporal decoding analyses time-locked to the onset of the probe do not reveal any significant temporal cluster, suggesting that negation does not invert the representation of the adjective to that of its antonym (e.g., “bad” to “good”) after the presentation of the probe number. For all panels: AUC = area under the receiver operating characteristic curve, chance = 0.5 (black horizontal dashed line); black vertical dashed lines indicate the presentation onset of the adjective in **A** and the probe in **B**; each line and shading represent participants' mean ± SEM.

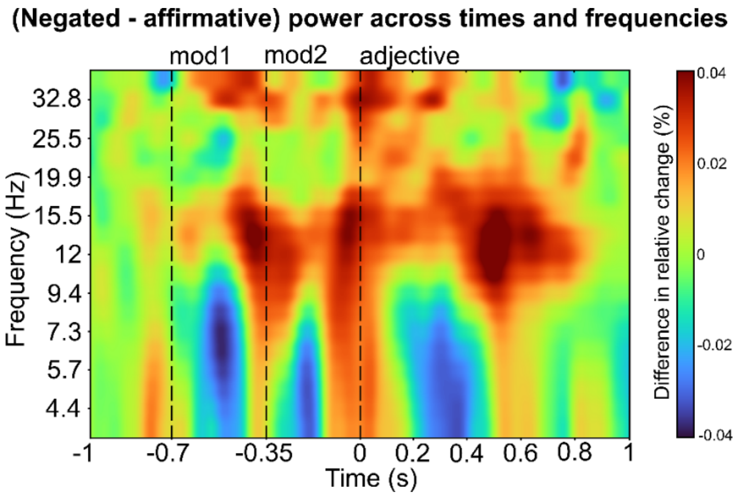


Fig. S5. Differences between negated and affirmative phrases across time and frequencies. Time-frequency spectrum of the differences between negated and affirmative phrases averaged across all sensors and all participants. Frequencies are between 3.9 and 37.2 Hz, logarithmically spaced. Black vertical dashed lines indicate the presentation onset of each word: modifier1, modifier2 and adjective; colors indicate % differences in change relative to a baseline of -300 to -100 ms from the onset of word 1 (modifier1).

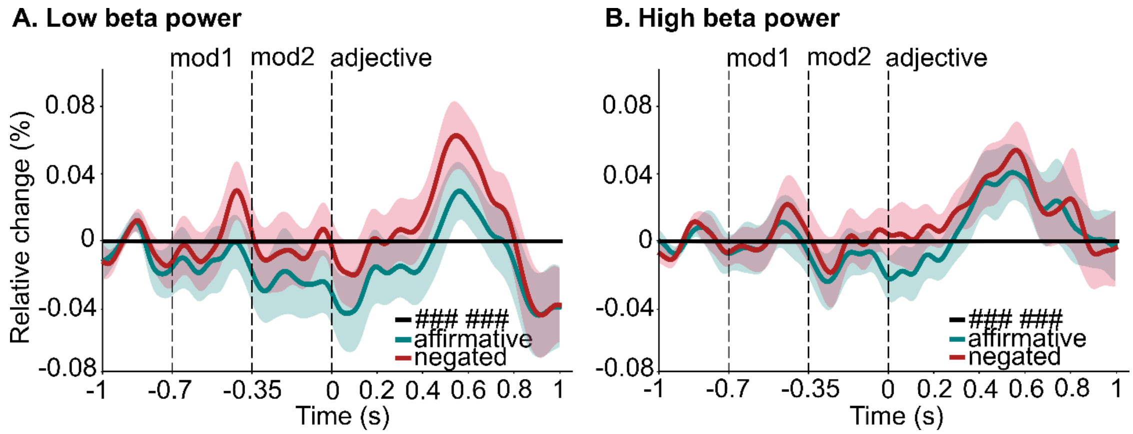


Fig. S6. Low- and high-beta power for negated and affirmative phrases across time. The mean beta power for the no modifier condition was subtracted from the mean beta power of affirmative and negated phrases, separately for low-beta (12-20 Hz, (**A**)) and high-beta (21-30 Hz, (**B**)). The horizontal solid black line represents the no modifier condition (i.e., ### ###) after subtraction (thus = 0), and the green and red lines represent beta power over time for affirmative and negated phrases, respectively. Relative change (%) was obtained by subtracting the mean of baseline values (-300 to -100 ms from the onset of word1) and dividing by the mean of baseline values. Black vertical dashed lines indicate the presentation onset of each word: modifier1, modifier2 and adjective; each line and shading represent participants' mean ± SEM.