Tracking the online construction of linguistic meaning through negation

Arianna Zuanazzi[1], Pablo Ripollés[1,2,3], Wy Ming Lin[4], Laura Gwilliams[5], Jean-Rémi King[1,6,*], David Poeppel[1,3,7,8,*]

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 Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany.

8 Ernst Strüngmann Institute for Neuroscience, Frankfurt, Germany.

*Equal contribution

Corresponding author: Arianna Zuanazzi, az1864@nyu.edu

Abstract

Combining words and composing meanings is at the basis of human language. However, how the brain constructs meaning online is not well understood. We develop an approach to address this puzzle that exploits the ubiquitous operation of negation. Although negation has been extensively debated among philosophers, psychologists, logicians, and linguists, the cognitive mechanisms supporting this elementary function remain largely unknown. Here, we track the compositional effects of negation ("not") on scalar adjectives (e.g., "good") in parametrically designed behavioral and magnetoencephalography (MEG) experiments. The behavioral data show that participants first interpret negated adjectives as affirmative and then modify their interpretation towards the opposite meaning. Decoding analyses of neural activity further reveal that negation does not invert the representation of adjectives but rather weakens their representation, as soon as lexical-semantic representations are encoded in the brain. This suppression mechanism is supported by desynchronization of beta-band activity elicited in sensory-motor areas during the processing of the adjective.

Introduction

"I once undertook a fairly systematic interrogation on the matter among intelligent acquaintances […] and they were practically unanimous in their testimony that they had never encountered a negative fact and that every case of knowledge expressed through a negative proposition was in reality of a positive nature, in a fashion which they were unable to comprehend." (Demos, 1917).

A hallmark of language processing is that we combine elements of the stored inventory - informally speaking, words - and thereby generate composed meanings that derive in systematic ways from the 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 comprehensive, mechanistic understanding of the underlying processes requires characterization of the operations underpinning syntactic structure building. Likewise, it is necessary to understand how meaning representations themselves are constructed in real time. There has been steady progress and productive debate on syntactic structure building (e.g., Ding et al., 2016; Fedorenko et al., 2016; Martin and Doumas, 2017;

Matchin and Hickok, 2020; Oseki and Marantz, 2020; Pallier et al., 2011). In contrast, how meaning composition is accomplished is less widely investigated, despite its foundational role in language comprehension and production. In the experimental approach pursued here, we build on the existing literature that focuses on precisely controlled 'minimal' linguistic environments (e.g., Bemis and Pylkkänen, 2011, 2013) to study the phenomenon of composition. We deploy a new and simple parametric experimental paradigm that capitalizes on the powerful role that *negation* plays in the construction of meaning. Using a combination of 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 bad") phrases. The results identify models and mechanisms of how negation, as a window into semantic composition, operates in real time.

Two aspects of negation are intuitive. First, although negation is typically understood correctly by interlocuters, the representation of negated concepts can be more difficult to grasp. Second, negation entertains some relation with the affirmative counterpart (e.g., "not good" ↔ "bad"). This latter aspect, and more generally the function of negation in natural language, has been a matter of longstanding debate among philosophers, psychologists, logicians, and linguists (Horn, 1989), and has been more recently studied in the context of artificial intelligence systems (e.g., Ettinger, 2020). Broadly speaking, negation is defined as the operation that reverses the meaning of a word, phrase, or sentence into its opposite meaning. In spite of its rich intellectual history and relevance (interpreting negation was, famously, an important point of debate between Bertrand Russell and Ludwig Wittgenstein), comparatively little research investigates cognitive and neural mechanisms of negation. Previous behavioral research shows that negated phrases/sentences are processed with more difficulty (i.e., slower and with more errors) than the affirmative counterparts, suggesting an asymmetry between negated and affirmative representations (e.g., Dale and Duran, 2011; Darley et al., 2020; Dudschig and Kaup, 2018; Just and Carpenter, 1971; Kaup et al., 2007; Kaup and Dudschig, 2020b). This asymmetry has motivated two fundamental questions in prior literature: (1) *how* does negation operate? and (2) *when* does negation operate?

Studies addressing the first question (i.e., *how*) have suggested that negation implies suppression: negation is argued to reduce the extent of available information (e.g., the availability of the information contained in "cookies" in the sentence "Mary bakes bread but no cookies", Kaup, 2001; Kaup and Zwaan, 2003; MacDonald and Just, 1989). Evidence that negation triggers

inhibitory processes is further provided by neuroimaging research investigating the neural dynamics of negation within the context of action representation. In action-related sentences such as "I do not close the door", negation is associated with decreased activity in brain regions within the action-representation network, such as premotor and primary motor cortices (i.e., *motor inhibition*, Bartoli et al., 2013; Djokic et al., 2019; Gallese and Lakoff, 2005; Papeo et al., 2016; Tettamanti et al., 2008; Tomasino et al., 2010). Furthermore, inhibitory control mechanisms (evidenced through theta/beta power increase) have been observed not only for putatively motor-related actions (e.g., "cut") but also for mental actions (e.g., "wish"), advancing the idea that negation could operate through a general-purpose response inhibition system (Beltrán et al., 2019; Beltrán et al., 2021; De Vega et al., 2016).

While previous neuroimaging studies have mainly focused on how negation affects action representation, psycholinguistic research has shown that scalar adjectives (e.g., "bad-good", "open-close", "full-empty") offer valuable insight into the dynamics of negation. These studies provide behavioral evidence showing that negation can either *eliminate* the negated concept and convey the opposite meaning (e.g., "not good" = "bad") or *mitigate* the meaning of its antonym along a semantic continuum (e.g., "not good" = "neither good nor bad" or "average"; Bianchi, 2011; Colston, 1999; Fraenkel and Schul, 2008; Giora, 2006). Thus, this system of polar opposites generated by scalar adjectives provides an especially useful linguistic framework to investigate changes in representation along the semantic scale (e.g., "bad" to "good"), as a function of negation (e.g., "bad" vs. "not good"), as well as offering a linguistic environment to address more traditional questions about compositionality (e.g., "good" vs "not + good" vs "really + not + good").

A second set of studies addressed the question of *when* negation operates. Traditional models suggest that understanding a negated phrase or sentence entails two steps: first, the affirmative representation is generated/simulated (e.g., "an open door" in "the door is *not* open"); second, the affirmative representation is suppressed/revised (e.g., "a closed door") (*two-step models*, Carpenter and Just, 1975; Clark and Chase, 1972; Kaup et al., 2006, 2007; Lüdtke et al., 2008; for a review, see Kaup and Dudschig, 2020a; Papeo and de Vega, 2020). In contrast, one-step accounts suggest that negation incrementally fuses with the representation of the words or phrase it modifies during one single step, affecting its representation already at the initial stages of semantic processing (i.e., *fusion models*; Horn, 1989; Lyons, 1995; Mayo et al., 2004; Papeo et al., 2016).

Here we capitalize on the semantic continuum offered by scalar adjectives to investigate how and when negation changes the representation of abstract concepts (e.g., “bad” vs “good”). First, we track semantic composition over time in a behavioral study deployed online. Next, we use magnetoencephalography (MEG) and a decoding approach to track the evolution of neural representations of target adjectives in affirmative and negated phrases. In particular, we test four salient hypotheses: (1) negation does not change the representation of the adjective (e.g., “not good” = “good”), (2) negation reduces the scalar value of the adjective (e.g., “not good” < “good”), (3) negation inverts the representation of the adjective (e.g., “not good” = “bad”), and (4) negation changes the representation of the adjective to another representation, outside the semantic scale defined by the two polar adjectives (e.g., “not good” = “unacceptable”). The combined behavioral and neurophysiological data adjudicate among these hypotheses and identify potential mechanisms (*how*) and temporal dynamics (*when*) that underlie how negation functions in online meaning composition.

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 and a scalar adjective (e.g., “really really good”, “really not quiet”, “not ### fast”. The adjectives were grouped into *low* antonyms (e.g., “bad”, “ugly”, “sad”, “cold”, “slow”, and “small”) and *high* antonyms (e.g., “good”, “beautiful”, “happy”, “hot”, “fast”, and “big”). The sequence of dashes was used to indicate the absence of a modifier. See Fig. 1A and supplementary Table S1 for a comprehensive list of the linguistic stimuli. On every trial, participants rated the overall meaning of each phrase on a scale defined by each antonyms pair (e.g., from “really really bad” to “really really good”, Fig. 1A). We analyzed continuous mouse trajectories and reaction times (RTs).

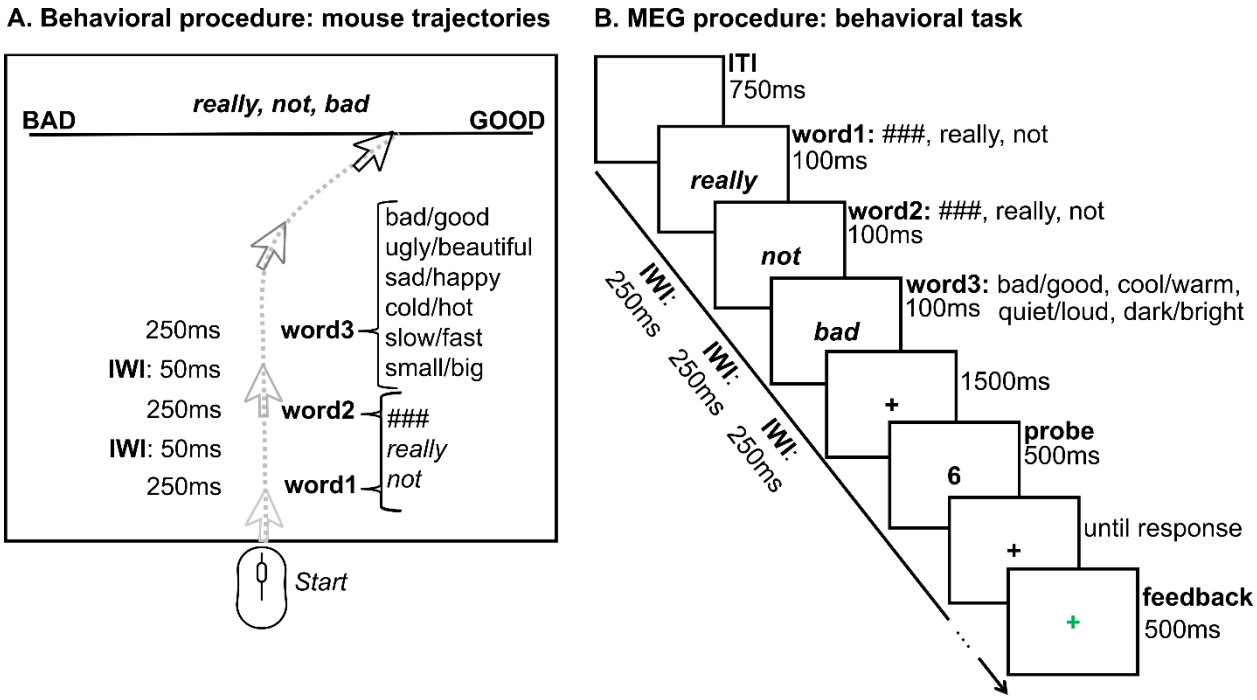


Figure 1. Experimental procedures.

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 from e.g., "really really bad" to "really really good". Each trial consisted of combinations of "###", "really", and "not" in word positions 1 and 2, and an adjective representing the low or high pole across six possible scalar dimensions. For each trial, we collected continuous mouse trajectories throughout the entire trial and reaction times. "###': no modifier. IWI: inter-word-interval. **B.** MEG procedure: participants read affirmative or negated adjective phrases (e.g., "really really good", ""### not bad") 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. "###': no modifier. IWI: inter-word-interval.

Reaction times: To evaluate the specific effect of antonym and of negation, 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 antonym ($F(1,77) = 60.83$, $p < 0.001$, $\eta_p^2 = 0.44$) 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 antonym and negation was observed ($p > 0.05$). These results indicate 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”). These results strengthen previous behavioral data showing that negation is associated with increased processing difficulty (e.g., Darley et al., 2020; Dudschig and Kaup, 2018). 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 natural language task.

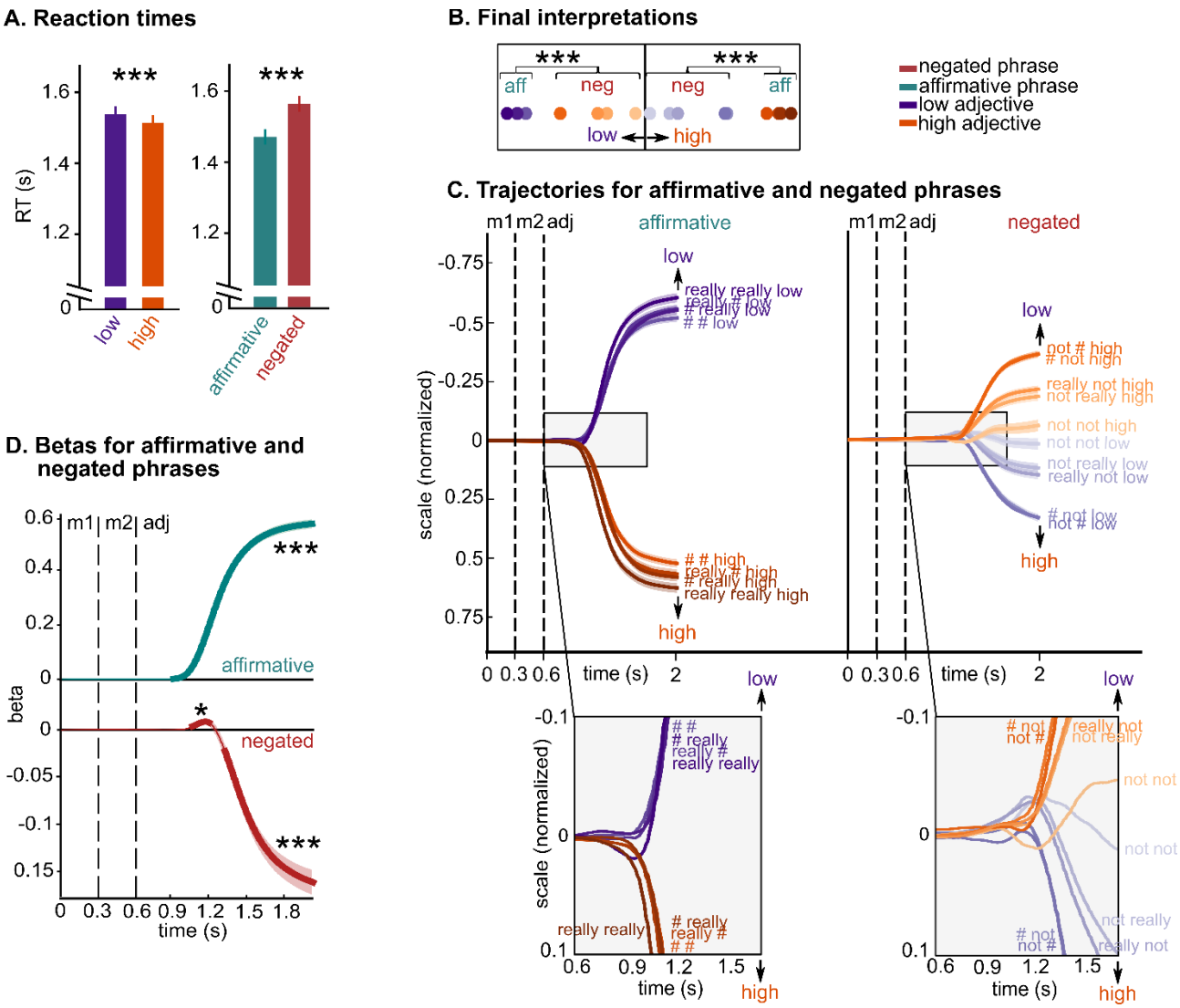


Figure 2. Behavioral results of 78 participants.

A. Reaction times results for the online behavioral study. **B.** Final interpretations of each phrase represented by filled dots (purple = low, orange = high), averaged across adjective pairs and subjects. These results show that the final rating of negated adjectives never overlaps with that of their antonyms (“not low” < “high”). **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. The zoomed-in panels at the bottom show that mouse trajectories of affirmative phrases branched towards the adjective's side of the scale and remained on that side until the final interpretation; conversely, the trajectories of negated phrases first deviated towards the side of the adjective and then towards the side of the antonym. This result is confirmed by linear models fitted to the data at each timepoint (**D**). **D**. Beta values (representing 78 participants) over time, separately for affirmative and negated phrases. 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; *** $p < 0.001$; * $p < 0.05$.

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 all dimensions, see supplementary Fig.S1 for each dimension separately).

To investigate how negation changes the interpretation of a scalar adjective, we performed a 2 (*antonym*: low vs high) x 2 (*negation*: negated vs affirmative) repeated-measures ANOVA for participants' final responses (filled circles in Fig.2B), which revealed a significant main effect of antonym ($F(1,77) = 338.57$, $p < 0.001$, $\eta_p^2 = 0.83$), a significant main effect of negation ($F(1,77) = 65.50$, $p < 0.001$, $\eta_p^2 = 0.46$), and a significant interaction between antonym and negation ($F(1,77) = 1346.07$, $p < 0.001$, $\eta_p^2 = 0.95$). Post-hoc tests showed 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, $p_{holm} < 0.001$). Results also showed that the final rating of negated phrases was significantly more variable (measured as standard deviation) 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 changes the final interpretation of the adjectives towards the opposite side of the scale never to a degree that overlaps with the antonym representation.

Second, we investigated the temporal dynamics of changes in adjective representation as a function of negation (i.e., from the presentation of word 1 to the final interpretation; lines in Fig.2C). 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 of Fig.2C), the trajectories of negated phrases first deviated towards the side of the adjective and then towards the side of the antonym (i.e., "not low" was first interpreted towards "low" and then towards "high", lines in the right gray zoomed-in panel of Fig.2C). To quantify the degree of deviation towards each side of the scale, we performed regression analyses

with antonym as the predictor and mouse trajectories as the dependent variable. The results confirmed this observation, showing that (1) in affirmative phrases, betas were positive starting from 300 ms from the adjective onset ($p < 0.001$, green line in Fig.2D); and that (2) in negated phrases, betas were positive between 450 and 580 ms from the adjective onset ($p = 0.04$), and only became negative from 700 ms from the adjective onset ($p < 0.001$, i.e., red line in Fig.2D).

Taken together, these results suggest that participants initially interpreted negated phrases as affirmative phrases (e.g., "not good" as "good") and then modified their interpretation towards the opposite meaning (e.g., "bad").

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

In Experiment 2 (in-lab MEG experiment, N = 26), participants read adjective phrases comprising one or two modifiers and scalar adjectives across different dimensions (e.g., "really really good", "really not quiet", "not ### dark'). The adjectives were grouped into *low* antonyms (e.g., "bad", "cool", "quiet", "dark") and *high* antonyms (e.g., "good", "warm", "loud", "bright"). The sequence of dashes was used to indicate the absence of a modifier. See Fig. 1B and supplementary Table S2 for a 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).

Similarly to the result of Experiment 1, behavioral data of Experiment 2 demonstrate that 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: $F(1,25) = 8.03$, $p = 0.009$, $\eta_p^2 = 0.24$).

Our MEG analysis comprises four steps: (1) we first identify the temporal and spatial correlates of simple word representation (i.e., the words "really", "not" in the modifier position, and each pair of scalar adjectives in the head position); (2) we turn to the representation of adjectives and negation; (3) we then ask how negation operates specifically on the representation of the adjectives, by teasing apart four possible mechanisms (i.e., *no effect*, *reduction*, *inversion*, *change*); (4) finally, we explore changes in beta power, as a function of negation.

Temporal and spatial decoding of single word processing

Fig.3A depicts the across trials grand average of preprocessed MEG data (see Methods for preprocessing details). We performed multivariate decoding analyses on these preprocessed MEG data. First, we evaluated single word representation, to ensure that our decoding approach is sensitive to visual/lexical-semantic differences between words. We performed a temporal decoding analysis (see Methods) where, for the first and second modifier separately and at each time sample, we estimated the ability of the decoder to discriminate "really" vs "not" from all MEG sensors. Results in Fig.3B show that the decoding of the modifier was 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 onset of the second modifier (light gray areas, $p < 0.001$).

Similarly, to evaluate whether the brain encodes representational differences between each pair of antonyms (including word form and lexical-semantic differences), we performed a temporal decoding analysis (see Methods) for pairs of adjectives in each of the four dimensions (*quality*, *temperature*, *loudness,* and *brightness*). Results in Fig.3C show that pairs of antonyms were similarly decodable between 90 and 410 ms from the onset of the adjective (*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, $p < 0.001$).

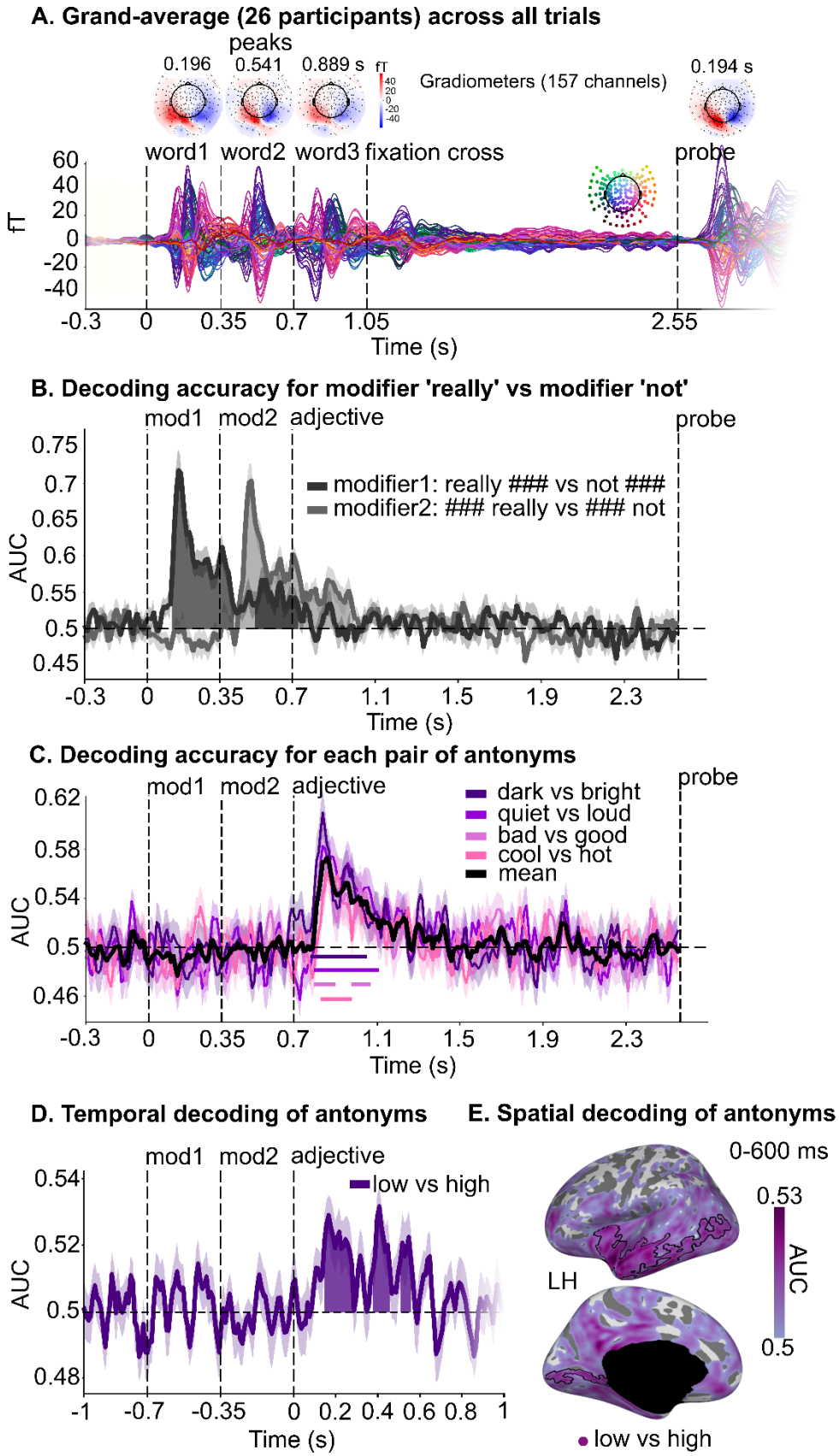

A. Grand-average (26 participants) across all trials
peaks
0.196
0.541
0.889 s
fT
Gradiometers (157 channels)
0.194 s
word1
word2
word3
fixation cross
probe
fT
Time (s)
B. Decoding accuracy for modifier 'really' vs modifier 'not'
mod1
mod2
adjective
probe
modifier1: really ### vs not ###
modifier2: ### really vs ### not
AUC
Time (s)
C. Decoding accuracy for each pair of antonyms
mod1
mod2
adjective
probe
dark vs bright
quiet vs loud
bad vs good
cool vs hot
mean
AUC
Time (s)
D. Temporal decoding of antonyms
mod1
mod2
adjective
low vs high
AUC
Time (s)
E. Spatial decoding of antonyms
0-600 ms
0.53
AUC
0.5
LH
low vs high

Figure 3. Evoked activity, temporal and spatial decoding of modifiers and adjectives.

A. The grand-average across all trials reveals evoked activity in occipito-temporal cortices (i.e., letter string processing) at each word presentation time. **B.** Decoding accuracy for the modifier identity ("really" vs "not"): this decoding analysis was performed only including the symmetrical phrases "really" and "not" combined with the placeholder "###", and was time-locked at each modifier presentation time. Significant windows are indicated by shaded gray areas. **C.** Decoding accuracy for antonyms in each dimension; the mean across adjective pairs is represented as a solid black line; significant windows are indicated by horizontal solid lines. **D.** Decoding accuracy (purple line) of lexical-semantic differences between antonyms over time (i.e., after pooling over the four dimensions); significant time windows are indicated by purple areas; **E.** Decoding accuracy (shades of purple) for antonyms over brain sources (after pooling over the four dimensions), between 0 and 600 ms from the onset of the adjective. Significant spatial clusters are indicated by a black contour. 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: modifier 1, modifier 2 and adjective; each line and shading represent participants' mean ± SEM; LH = left hemisphere.

Temporal decoding of antonyms and negation

Here, we move beyond single word representation, to evaluate lexical-semantic differences between antonyms of the same scale. We entered *low* ("bad", "cool", "quiet" and "dark") and *high* ("good", "warm", "loud" and "bright") antonyms in the same model. The temporal decoding analysis (see Methods) revealed that the decoding accuracy of antonyms was significantly above chance in three time-windows between 140 and 560 ms from the onset of the adjective (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.3D). No differences in lexical-semantic representation were observed in later time windows (i.e., between 500 and 1000 ms from the onset of the adjective). The spatial decoding analysis (limited to the first 600 ms from the onset of the adjective) showed that decoding accuracy was significantly above chance in a widespread left-lateralized brain network, encompassing the anterior portion of the superior temporal lobe, the middle and inferior temporal lobe, the lateral occipital lobe, the insula, the pericalcarine and the lingual gyrus (purple areas in Fig.3E, significant clusters are indicated by a black contour: left temporal lobe cluster, $p = 0.002$; left calcarine cluster, $p = 0.032$; no significant cluster was found in the right hemisphere).

Next, we turn to lexical-semantic representations of negation over time. We performed a temporal decoding analysis (see Methods) for phrases containing negation vs phrases not containing negation, separately for phrases with one and two modifiers (to account for phrase

complexity; see supplementary Table S2 for a list of all trials). Results revealed that, for phrases with one modifier, the decoding of negation is significantly higher than chance throughout word 1 (-580 to -500 ms from the onset of the adjective, $p = 0.005$), then again throughout word 2 (-470 to 0, $p < 0.001$). 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 the onset of the adjective ($p = 0.018$; dark red line and shading in Fig.4). Similarly, for phrases with two modifiers, the decoding of negation is significantly higher than chance throughout word 1 (-580 to -410 ms from the onset of the adjective, $p = 0.002$), then again throughout word 2 (-400 to 0, $p < 0.001$). After the presentation of the adjective, negation decodability is significantly above chance for a longer time window 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.4). The same analysis time-locked to the onset of the probe shows that negation is once again significantly decodable between 230 and 930 ms (supplementary Fig.S3).

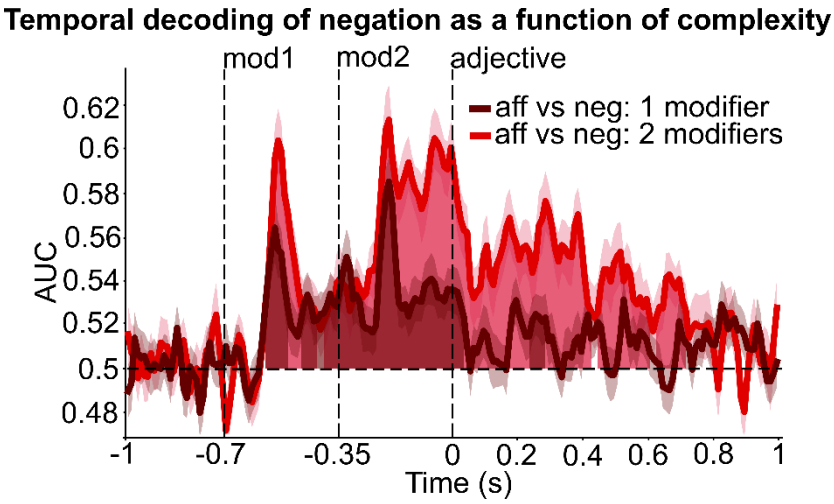


Figure 4. Temporal decoding of negation as a function of number of modifiers (i.e., complexity).
Decoding accuracy of negation over time, as a function of the number of modifiers (1 modifier: dark red line and shading; 2 modifiers: light red line and shading). Significant time windows are indicated by dark red (1 modifier) and light red (2 modifiers) areas. 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; aff = affirmative, neg = negated; mods = modifiers; each line and shading represent participants' mean ± SEM.

Overall, these results suggest that the brain encodes negation every time a "not" is presented, and maintains this information available up to 720 ms after the adjective onset. Further, they show that the duration of negation maintenance is boosted by the presence of a second modifier, suggesting that combinatorics efforts increase with the number of modifiers in the phrase.

Effect of negation on antonym representation over time

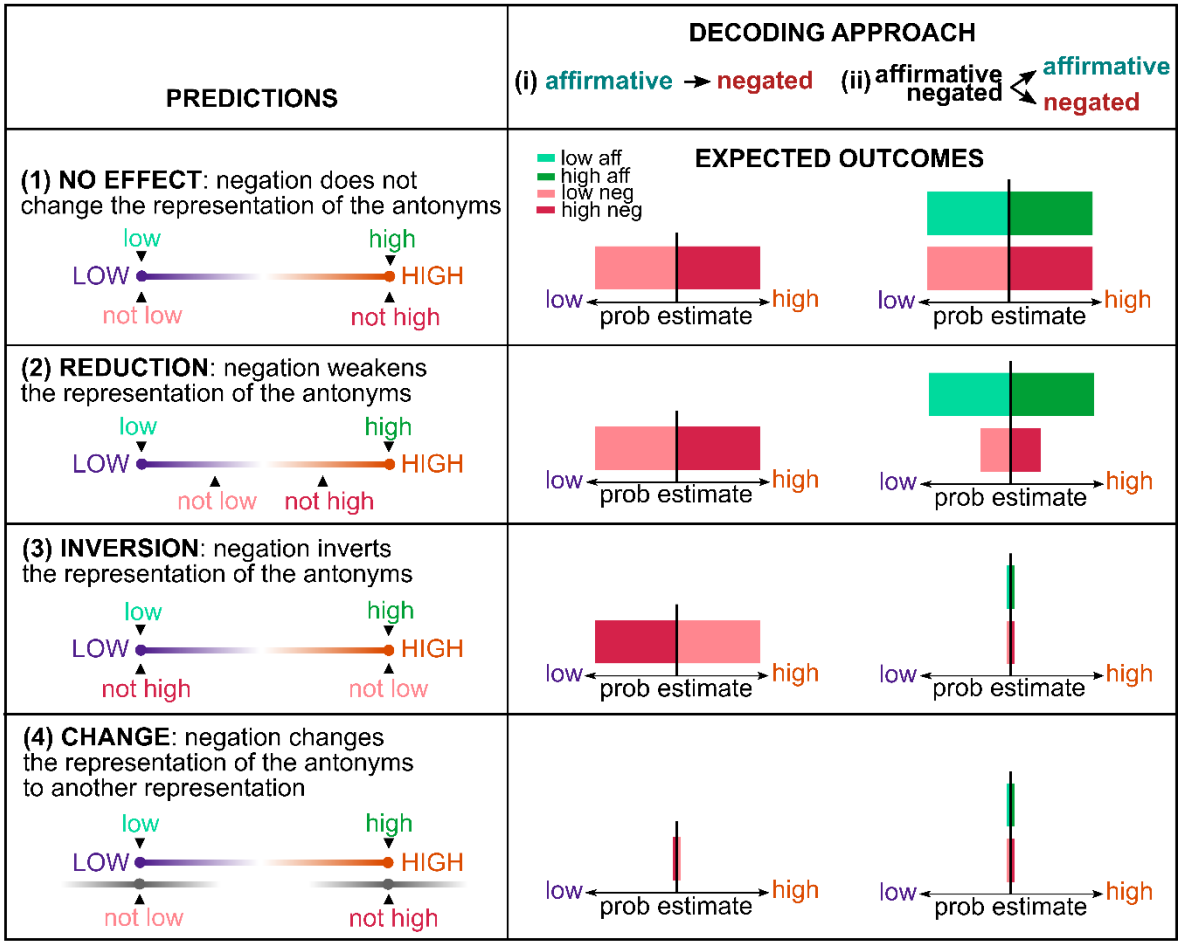
The temporal decoding of the adjective and of negation demonstrate that the brain maintains the representation of the operator "not" available throughout the presentation of the adjective.

Here, we ask how negation operates on the representation of the antonyms at the neural level, leveraging theoretical accounts of negation and two complementary decoding approaches. We test four hypotheses (left column in Fig.5A): (1) *No effect of negation*: negation does not change the representation of the adjective (i.e., "not low" = "low", "not high" = "high"); (2) *Reduction*: negation weakens the representation of the adjective (i.e., "not low" < "low", "not high" < "high"); (3) *Inversion*: negation inverts the representation of the adjective (i.e., "not low" = "high", "not high" = "low"); (4) *Change*: negation changes the representation of the adjective to another representation, outside the semantic scale defined by the two antonyms (e.g., "not low" = "fair", "not high" = "average").

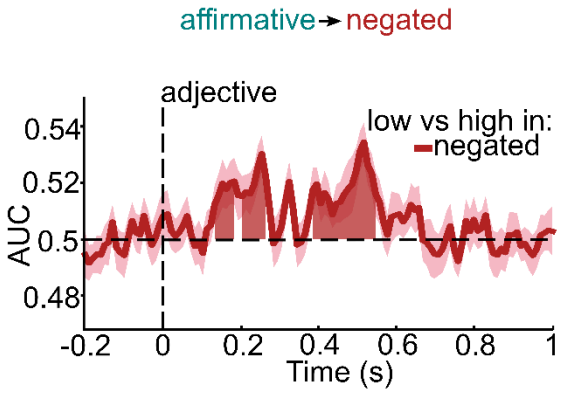
To arbitrate between these four predictions, we performed two sets of decoding analyses: (i) we computed the accuracy with which estimators trained on low vs high antonyms in affirmative phrases (e.g., "really really bad" vs "really really good") generalize to the representation of low vs high antonyms in negated phrases (e.g., "really not bad" vs "really not good") at each time sample time-locked to the adjective onset (see Methods). Expected probability estimates (i.e., the averaged class probabilities for *low* and *high* classes) are depicted as light and dark red bars in the right column in Fig.5A. For our results to support predictions (1) *No effect* or (2) *Reduction*, we should observe a significant above chance AUC, and probability estimates for high and low adjectives above the computed chance level in the direction of the respective class, indicating that the initial representation of the adjectives in negated phrases is similar to that in affirmative phrases. Conversely, for our results to support prediction (3) *Inversion*, we should observe a significant below chance AUC, and probability estimates for low and high adjectives above the computed chance level in the direction of the opposite class (i.e., swapped), as adjective

representations would be systematically inverted in negated phrases. Finally, we should observe at chance AUC and probability estimates in the case of (4) *Change*, where adjective representations in negated phrases are not predictable from the corresponding representations in affirmative phrases. To be able to disentangle predictions (1) *No effect* from (2) *Reduction*, we (ii) 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 the remaining 10% of the trials, for affirmative and negated phrases separately (see Methods). This approach allows for direct comparison between AUC and probability estimates in affirmative and negated phrases. For our results to support prediction (1) *No effect*, we should observe comparable AUC and probability estimates in affirmative and negated phrases, suggesting that negation does not change the representation of the adjective. Conversely, in support of prediction (2) *Reduction*, we should observe reduced AUC and probability estimates for negated than affirmative phrases, suggesting less robust differences between low and high antonyms in negated phrases. The outcome of prediction (3) *Inversion* and (4) *Change* would be at chance AUC and probability estimates (as the model is trained on different representations within the same class). Fig.5A illustrates a summary of our predictions and decoding approaches.

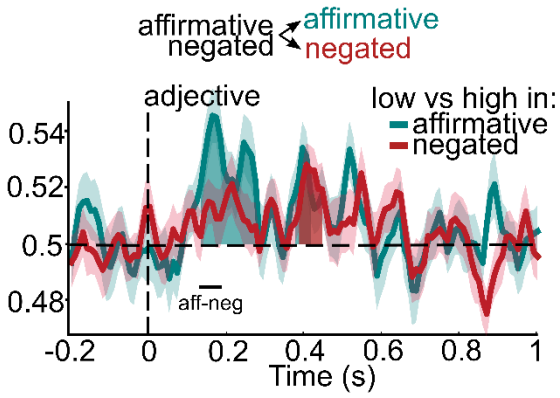
A. Expected outcomes for the effect of negation on the representation of antonyms



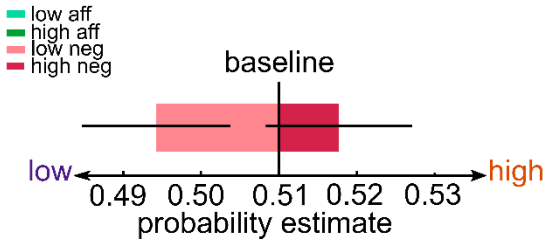
B. Results of decoding approach (i)



C. Results of decoding approach (ii)



D. Significant clusters (averaged)



E. Significant clusters (averaged)

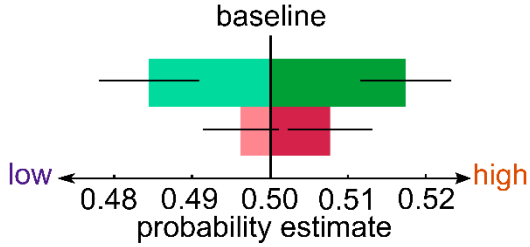


Figure 5. Expected outcomes and results of the effect of negation on antonym representation.

A. We arbitrated between four possible effects of negation on the representation of the adjective: (1) *No effect*, (2) *Reduction*, (3) *Inversion*, (4) *Change*. Note that we depicted (1) *No effect*, and (3) *Inversion* on a continuous low-high representation scale, but a discrete scale would predict the same outcome. We performed two sets of decoding analyses: (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. Significant time windows of the difference between affirmative and negated phrases are 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**. **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 (black vertical line) was computed by averaging probability estimates of the respective baseline (note that the baseline can differ from 0.5 as a function of the number of trials in each class in the training set). For all panels: AUC = area under the receiver operating characteristic curve, chance = 0.5 (black horizontal dashed line); the black vertical dashed line indicates the presentation onset of the adjective; purple = low adjectives, orange = high adjectives; green = affirmative phrases, red = negated phrases; each line and shading (or error bar) represent participants' mean ± SEM.

Temporal decoding analysis (i) revealed 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 the onset of the adjective (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 (chance level was computed by averaging probability estimates of the respective baseline). These results only support the predictions (1) *No effect* and (2) *Reduction*, thus invalidating predictions (3) *Inversion* and (4) *Change*. Supplementary Fig.S2 illustrates a different approach where phrases such as “not good” and “really bad” were both labeled as “bad”. This approach similarly leads to the exclusion of prediction *(3) Inversion*.

Temporal decoding analysis (ii) revealed 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 the onset of the adjective. Conversely, decoding accuracy for negated phrases is significantly above chance only between 380 and 450 ms after the onset of the adjective ($p = 0.004$). Moreover, negated phrases are associated with significantly lower decoding accuracy than affirmative phrases in the time window between 130 and 190 ms from the onset of the adjective (black horizontal line in Fig.5C; $p = 0.040$). Fig.5E depicts the probability estimates averaged over this 130-190 ms significant time-window for low and high antonyms, separately in affirmative and negated phrases (chance level was computed by averaging probability estimates of the respective baseline).

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 the adjective onset (i.e., within the time window of semantic lexical-semantic processing, see Fig.3D) provide evidence in support of the (2) *Reduction* prediction, wherein negation weakens the representation of the adjectives.

Changes in beta power as a function of negation

The previous results tease apart different possible effects of negation on the representation of antonyms and demonstrate that negation does not invert or change the representation of the adjectives but reduces the decodability of low vs high antonyms, significantly for about 60 ms from the onset of the adjective. The availability of negation upon the presentation of the adjective and the reduced decoding accuracy for antonyms in negated phrases raises the question of whether negation operates through inhibitory mechanisms, as suggested by previous research employing action-related verbal material (Beltrán et al., 2019; De Vega et al., 2016). We performed time-frequency analyses focusing on beta power (including low beta: 12 to 20 Hz, and high beta: 20 to 30 Hz, Weiss and Mueller, 2012), which has been previously associated to inhibitory control (e.g., Wagner et al., 2018, see supplementary Fig.S4 for comprehensive time-frequency results). We reasoned that, if negation operates through a general-purpose inhibitory system, we should observe increased beta power for negated than affirmative phrases in sensory-motor brain regions.

Our results confirmed this hypothesis, by showing significantly increased low-beta power for negated than affirmative phrases (from 229 to 350 ms from the onset of modifier1: $p = 0.036$; from 326 to 690 ms from the onset of the adjective: $p = 0.012$; red line in Fig.6A). We also observed

significantly increased high-beta power for negated than affirmative phrases (from 98 to 271 ms from the onset of the adjective: $p = 0.044$; yellow line in Fig.6A). Supplementary Fig.S5 further shows low and high beta power separately for negated and affirmative phrases, compared to phrases with no modifier.

At the source level, our whole-brain analysis showed significantly higher low-beta power for negated than affirmative phrases in the left precentral, postcentral and paracentral gyrus ($p = 0.012$; between 326 and 690 ms from the onset of the adjective, red clusters in Fig.6C). No significant spatial clusters were found for high-beta power (however similar, albeit not significant, sensory-motor spatial patterns emerged, see yellow clusters in Fig.6B).

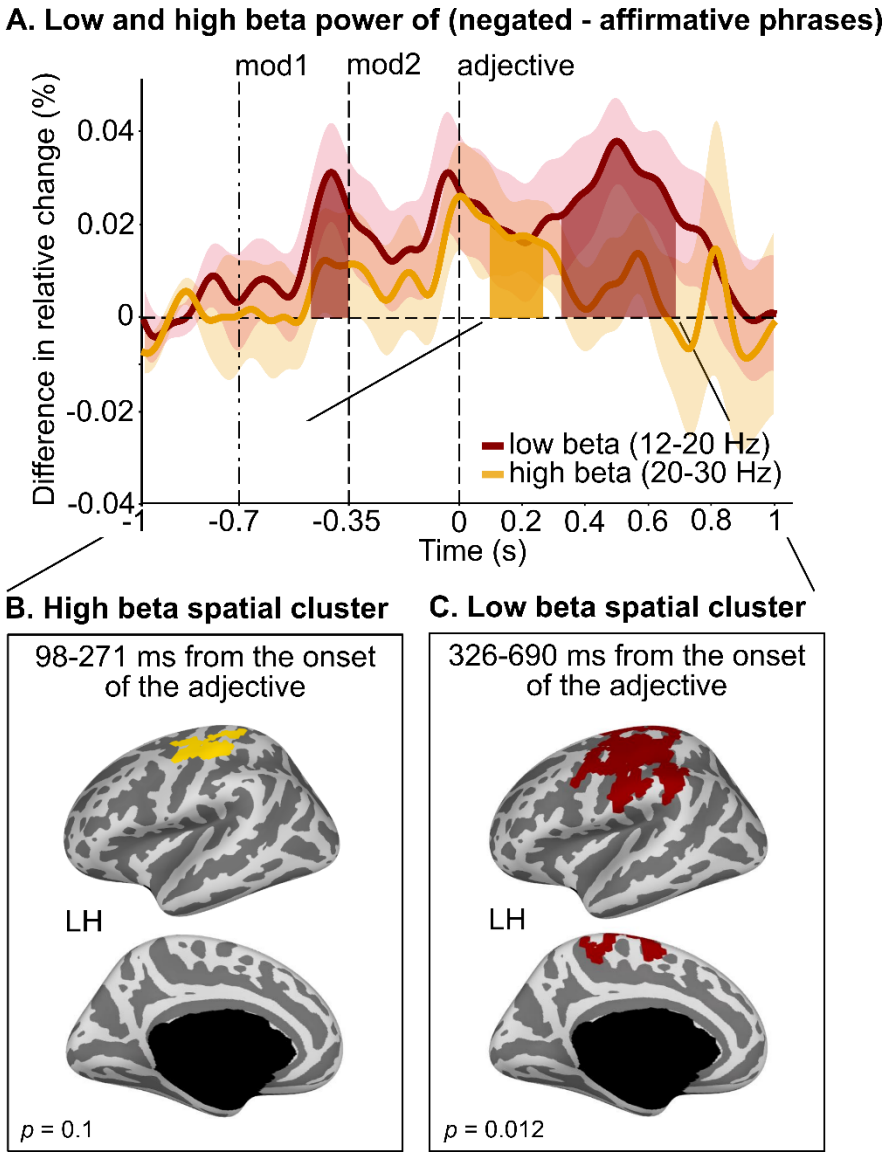


Figure 6. Differences in beta power over time between negated and affirmative phrases.

A. Differences in low (12-20 Hz, red) and high (20-30 Hz, yellow) beta power over time between negated and affirmative phrases. Overall, negated phrases showed increased 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**. (Non-significant) differences in high-beta power between negated and affirmative phrases (restricted between 97 and 271 ms from adjective onset). **C**. Significant differences in low-beta power between negated and affirmative phrases (restricted between 326 and 690 ms from adjective onset) were found in the left precentral, postcentral and paracentral gyrus (red cluster). No significant spatial clusters were found in the right hemisphere.

Discussion

In this behavioral and MEG study we tracked changes in lexical-semantic representation of scalar adjectives over time, as a function of negation. We leveraged psycholinguistics findings on scalar adjectives that offer a unique linguistic framework where meaning is represented on a continuum (Bianchi, 2011; Colston, 1999), time-resolved behavioral and neural data, and multivariate analysis methods (i.e., decoding). The neural correlates of negation have been mostly investigated in the context of action verbs (e.g., "write") or verbs that refer to mental states (e.g., "remember"; Alemanno et al., 2012; Aravena et al., 2012; Beltrán et al., 2019; De Vega et al., 2016; Foroni and Semin, 2013; Ghio et al., 2018; Liuzza et al., 2011; Papeo et al., 2016; Tettamanti et al., 2008; Tomasino et al., 2010; however see: Dudschig et al., 2018; Lüdtke et al., 2008; Nieuwland and Kuperberg, 2008; Palaz et al., 2020; Xiang et al., 2016). Our study employs minimal linguistic contexts to specifically tackle how negation operates on more abstract and non-action-related semantic representations.

Results of the continuous behavioral trajectories (Experiment 1) reveal that participants were slower to report the composed meaning of negated than affirmative phrases. This result is replicated in our MEG experiment (Experiment 2), where participants were not only slower in the *yes/no* task for negated phrases but were also less accurate. These data are consistent with previous behavioral studies showing that negation typically incurs increasing processing costs (e.g., Dale and Duran, 2011; Darley et al., 2020; Dudschig and Kaup, 2018; Kaup et al., 2007; Kaup and Dudschig, 2020; Just and Carpenter, 1971).

Complementing information provided by RTs, behavioral trajectories further allow to collect time-resolved data that reflect dynamic changes in representation (Darley et al., 2020; Dotan and Dehaene, 2013; Maldonado et al., 2019). Our mouse trajectories results reveal that participants initially interpreted negated phrases as affirmative phrases (e.g., "not good" was interpreted as "good", for ~130 ms) and then modified their interpretation towards the opposite meaning (e.g., "bad"). Furthermore, results of participants' final responses highlight that the interpretation of negated antonyms never overlapped with that of the affirmative counterpart on the semantic scale (e.g., "not good" < "bad"), suggesting that, even at the final stage of interpretation, negation does not completely invert the meaning of the adjective to that of its antonym (Bianchi, 2011; Colston, 1999; Fraenkel and Schul, 2008; Giora, 2006).

Informed by our behavioral data, we turned to MEG, which offers the opportunity to track neural changes in representation with high temporal and spatial resolution. We first identified the temporal and spatial correlates of lexical-semantic processing of adjectives and negation separately, which sets the basis for operationally characterizing the effect of negation on the representation of adjectives. Our temporal decoding analysis highlights a time window between 140 and 560 ms from the onset of the adjective when the brain encodes differences between *low* ("bad", "cool", "quiet", "dark") and *high* adjectives ("good", "warm", "loud", "bright"). This time window is consistent with studies investigating lexical-semantic processing in language comprehension. In reading tasks, access to lexical and semantic properties of words has been identified as early as 130–200 ms, up to ~550 ms from word onset (Hauk et al., 2006; Kutas and Federmeier, 2011; Lau et al., 2008; Pulvermüller et al., 2001, 2009; Teige et al., 2018). Note that in our data the earliest time of antonym decodability (i.e., ~140 ms) cannot reflect lower-level word form or lexical frequency differences, as we pooled over four adjectives for each *low* and *high* class, and we matched these classes for frequency. Conversely, word form differences between antonyms emerge at the single word level (i.e., within each dimension, e.g., *goodness*: "good" vs "bad"), starting as early as 90 ms from the adjective onset. Our spatial decoding analysis highlighted a left-lateralized occipito-temporal cluster encoding differences between *low* and *high* adjectives. These brain areas are part of the left-lateralized frontotemporal brain network that underpins semantic representation and composition, such as the inferior frontal gyrus, temporal cortices (including the anterior temporal lobe), angular gyrus, fusiform gyrus, and posterior cingulate gyrus (Bemis and Pylkkänen, 2011, 2013; Binder et al., 2009; Caucheteux et al., 2021;

Huth et al., 2016; Lau et al., 2013; Lambon Ralph et al., 2017; Noonan et al., 2013; Popham et al., 2021; Pulvermüller, 2013). Altogether, we show that lexical-semantic features of adjectives are encoded within a time-window and in brain networks that have been previously shown to underlie semantic processing across a large variety of stimulus types and sensory modalities (Berezutskaya et al., 2020; Kutas and Federmeier, 2011; Patterson et al., 2007; Vandenberghe et al., 1996).

In parallel to identifying the lexical-semantic representations of adjectives, we showed that negation is significantly decodable at each modifier's presentation time (reflecting word form encoding and semantic encoding), and its decodability remains sustained throughout the presentation of the adjective, up to ~700 ms (see Zhang and Pylkkänen, 2018 for a similar result at the sentence level). This pattern was mainly observed in phrases with two modifiers, likely reflecting syntactic and semantic compositional processes (Fedorenko et al., 2016; Pallier et al., 2011; Parrish and Pylkkänen, 2022). These results demonstrate that the lexical-semantic features of adjective and negation are processed in a parallel fashion, as opposed to serially (see Fyshe et al., 2019 for a similar pattern for adjective and noun in the context of adjective-noun phrases).

We then evaluated the effects of negation on adjective representation (i.e., *how* negation operates), by teasing apart the hypotheses that negation (1) does not change the representation of the scalar adjective (e.g., "not good" = "good", *No Change*), (2) weakens the representation of the scalar adjective (e.g., "not good" < "good", *Reduction*), (3) inverts the representation of the scalar adjective (e.g., "not good" = "bad", *Inversion*), or (4) changes the representation of the scalar adjective to another representation, outside the semantic scale defined by the two polar adjectives (e.g., "not good" = "unacceptable", *Change*). First, we demonstrated that, within the time window of adjective encoding, the representation of affirmative adjectives generalizes to that of negated adjectives. This result invalidates predictions (3) *Inversion* and (4) *Change*, and complements our behavioral data that show that negated adjectives are initially interpreted by participants as affirmative. Second, we demonstrated that the representation of adjectives in affirmative and negated phrases is not identical, but is weakened by negation. This result invalidates prediction (1) *No Effect* and supports prediction (2) *Reduction*, wherein negation reduces the scalar value of the adjective.

Previous research on negation has suggested that the opposite meaning of a scalar adjective would be more easily conveyed by the affirmative counterpart than by negation. For instance, if the speaker wants to convey the opposite meaning of "bad", they will more likely use "good" as

opposed to “not bad” (Grice, 1975; Horn, 1984); thus, negation would be purposely used (and understood) to convey a different or weakened meaning of the antonym (e.g., “not bad” = “a bit less bad than bad”). Our results bring evidence in support of this hypothesis. Similar conclusions are also reached by the behavioral study of Bianchi et al. (2011), who carefully characterized the shift produced by negation on adjectives across different polar dimensions. In their study, negated adjectives were never interpreted by participants as overlapping with the opposite end-pole, but rather along the semantic scale. This behavior has been observed especially for antonyms that are *contraries* (i.e., that lie on a continuum, e.g., “bad” and “good”) compared to *contradictories* (i.e., that create a dichotomy, e.g., “dead” and “alive”, Fraenkel and Schul, 2008). Here, we tested only adjective pairs that form contraries. Thus, different patterns of results could emerge in the case of contradictories that lack a middle point.

Turning to the question of *when* negation operates, we showed that the neural effects of negation are evident starting from early lexical-semantic processing times (i.e., from 130 ms from the onset of the adjective), providing new evidence in favor of one-step accounts, which suggest that negation operates as soon as semantic effects are detected. Similar results are reported by Papeo et al. (2016) who, using transcranial magnetic stimulation (TMS), showed that negation selectively affects the representation of action-related words already at 250 ms from the onset of the word, evidenced by increased motor inhibitory activity. While previous ERP studies have hypothesized that negation might be integrated with the meaning of words and phrases at a later time window (e.g., 550-1000 ms; Lüdtke et al., 2008), we did not decode adjective representations (or inverted representations) after 560 ms from the adjective onset. Future neuroimaging studies are needed to further elucidate how abstract representations are modulated by negation after the initial lexical-semantic processing has occurred.

The weakening effect of negation on the representation of adjectives points to the possibility that negation operates as a suppression mechanism, through general-purpose inhibitory systems (Beltrán et al., 2021). To address this question, we compared beta power modulations in affirmative and negated phrases. Beside motor processing, beta-power modulation (12-30 Hz) has been previously associated with multiple aspects of language processing (for a review, see Schaller et al., 2017; Weiss and Mueller, 2012). In particular, low-beta activity (i.e., ~12-20 Hz) is related with the processing of semantic information, evidenced in tasks employing semantic violations (Luo et al., 2010), contrasting concrete and abstract nouns (Weiss and Rappelsberger, 1996), open-

and closed-class words (Bastiaansen et al., 2005), words and pseudowords (Supp et al., 2004), and affirmative and negated sentences (Beltrán et al., 2019), among others. Thus, we evaluated differences between negated and affirmative phrases separately in the low- and high-beta bands. Our results show greater power for negated than affirmative phrases in both low- and high-beta bands. This difference is significant during the processing of the modifier and throughout the processing of the adjective until ~700 ms from the adjective onset, and is localized in left-lateralized sensory-motor areas. The timing and the spatial correlates of beta-power in relation to negation are in line with previous studies that examined the effect of negation on (mental and motor) action representation (Beltrán et al., 2021). Strikingly, our data demonstrates that negation recruits brain areas and neurophysiological mechanisms linked to suppression which are similar to that of response inhibition, however in the absence of action-related language material and by employing a task that does not require motor inhibition.

In conclusion, in this study we demonstrated that negation weakens (rather than inverts) the semantic representations of scalar adjectives, likely by recruiting response inhibition networks. Our study constitutes an important contribution to the understanding of composition as a foundational mechanism of language comprehension. By specifically tackling the construction of linguistic meaning through negation, and by employing time-resolved behavioral and neuroimaging methods and multivariate analyses, we were able to pinpoint subtle changes in semantic representation and tease apart different possible outcomes of semantic composition, above and beyond the sum of the individual word meanings.

Materials and Methods

Participants

Study 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 (an internal NYU platform for students' recruitment). All participants were native English speakers and self-reported normal hearing 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, Dotan and Dehaene, 2013; Pinheiro-Chagas et al., 2016; Darley et al., 2020) and was increased to account for the exclusion rate reported for online crowdsourcing experiments (Simcox and Fiez, 2014; Peer et al., 2017).

Experiment 2: MEG. 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 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 the 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, Lüdtke et al., 2008, De Vega et al., 2016; Beltrán et al., 2019), investigating semantic representation using MEG (22 to 25 participants, Bemis and Pylkkänen, 2011, 2013), or employing decoding methods with MEG data (17 to 20 participants, King et al., 2016, Gwilliams and King, 2020).

Stimuli, Design and Procedure.

Experiment 1: continuous mouse tracking.

Stimuli and Design. The linguistic stimulus set comprises 108 unique adjective phrases (for the complete list, see supplementary 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 *contraries* (i.e., adjectives that lie on a continuum, Fraenkel and Schul, 2008). Lexical characteristics of the antonyms were balanced according to the English Lexicon Project (Balota et al., 2007; 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 adverb "really" or "not". 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). Each dimension (e.g., temperature) 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. The online experiment was developed using oTree, a Python-based framework for the development of controlled experiments on online platforms (Chen et al., 2016). 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). The solid line represented the direction of the scale (*low to high* or *high to low*). 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. 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 (SOA: 50 ms). The "start" button was placed in the center portion of the bottom of the response space (i.e., in a neutral position). After each trial, participants were provided feedback: "incorrect" if the cursor's movement violated the warnings provided during the familiarization phase or 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.

Experiment 2: MEG.

Stimuli and Design. The linguistic stimulus set comprised 72 unique adjective phrases (for the complete list, see supplementary 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 (Balota et al., 2007; 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), for "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 adverb "really" or the operator "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). In this experiment, the condition with no modifiers ("### ###") was only employed as a baseline comparison in the time-frequency analysis and thus was excluded from all other analyses (to avoid possible differences in neural representation of phrases with and without syntactic/semantic

composition). Each dimension (e.g., temperature) 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 (SOA: 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 (Kleiner et al. 2007; 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 behavioral experiment was approximately 60 minutes.

Data analysis

Experiment 1: 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 subject were entered into 2 (*antonym*: low vs high) x 2 (*negation*: negated vs affirmative) repeated-measures ANOVA.

To evaluate differences in the final rating between antonyms in each scale, between negated and affirmative phrases, and their interactions, mean and standard deviation of the final responses of each subject were entered into a 2 (*antonym*: low vs high) x 2 (*negation*: negated vs affirmative) repeated-measures ANOVA. 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 subject, 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 Dotan and Dehaene, 2013). The dependent variable was the mouse coordinate along the scale (note that the scale, which was reversed in half of the trials, was swapped 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 the onset of the adjective (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 subject were entered into 2 (*antonym*: low vs high) x 2 (*negation*: negated vs affirmative) repeated-measures ANOVA (paired-samples t-test were performed to evaluate simple main effects of significant interactions).

The response time analysis was limited to correct trials. RTs within 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 subject 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 (Gramfort et al., 2013) 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 (Adachi et al., 2001). 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, heartbeats and external noise sources. MEG recordings were then epoched into epochs of -300 ms and 2550 ms around the onset of the first word 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 analysis). Note that, for visualization purposes, only 1700 ms from the onset of the first word 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 subjects, we manually scaled and co-registered the "fsaverage" brain to the participant's head-digitalized shape and fiducials (Gramfort et al., 2013; Andersen, 2018).

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: Dale et al., 2011). 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 (Gramfort et al., 2013) and Scikit-Learn packages (Pedregosa et al., 2011). 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. 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 subject 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), that fitted the linear estimator to 80% (or 90%, depending on the number of trials per class) 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, 3C, 3D, 4 and 5C (*decoding approach (ii*)). To investigate whether the representation of antonyms was comparable between affirmative and negated phrases, in a different set of analyses (*decoding approach (i)*, Fig.5B) we fitted the linear estimator to the epochs corresponding to affirmative phrases and generated predictions on the epochs corresponding to negated phrases. Accuracy and probability estimates for each class were then computed. For all decoding analyses, 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 the AUC across subjects were significantly higher than chance using a one-sample permutation cluster test, as implemented in MNE-python (10000 permutations, Maris and Oostenveld, 2007). We performed separate permutation cluster tests for the time windows: -700 to -350 ms from the onset of the adjective (i.e., word 1), -350 to 0 ms from the onset of the adjective (i.e., word 2), 0 to 500 ms from the onset of the adjective (i.e., time-window for early and late lexical-semantic processes; Kutas and Federmeier, 2011; Pulvermüller et al.,2009) and 500 to 1000 ms from the onset of the adjective (i.e., to account for potential later processes, Chen et al., 2016; Fyshe et al., 2019).

Spatial decoders. We fitted each estimator on each subject separately, across 600 time samples relative to the onset of the adjective (i.e., to include the three significant time windows obtained from the temporal decoding analysis in Fig. 3B), at each brain source separately, after morphing individual subject's source estimates to the ico-5 "fsaverage" common reference space. We employed a 5-fold stratified cross-validation, that 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 AUCs across subjects were significantly higher than chance using a one-sample permutation cluster test as implemented in MNE-python (10000 permutations; adjacency computed from the "fsaverage" brain; Maris and Oostenveld, 2007).

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 (20 to 30 Hz, Weiss and Mueller, 2012; Schaller et al., 2017) 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 subjects was significantly greater than 0, using a one-sample permutation cluster test as implemented in MNE-python (10000 permutations; Maris and Oostenveld, 2007). We performed separate permutation cluster tests for 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 (no significant differences were observed in analyses ran for later time-windows). We then computed the induced power in source space (method: dSPM and morphing individual subject'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 subjects was significantly greater than 0, using a one-sample permutation cluster test (10000 permutations; adjacency computed from the "fsaverage" brain; Maris and Oostenveld, 2007).

Acknowledgements

This work was supported by the Leon Levy Foundation (AZ), the European Union's Horizon 2020 research and innovation program under grant agreement No 660086 (JRK), the Bettencourt-Schueller Foundation (JRK), the Fondation Roger de Spoel-berch (JKR), the Philippe Foundation (JRK), and the FrontCog grant ANR-17-EURE-0017 (JRK).

Supplementary tables and figures:

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 for the time-frequency analysis.

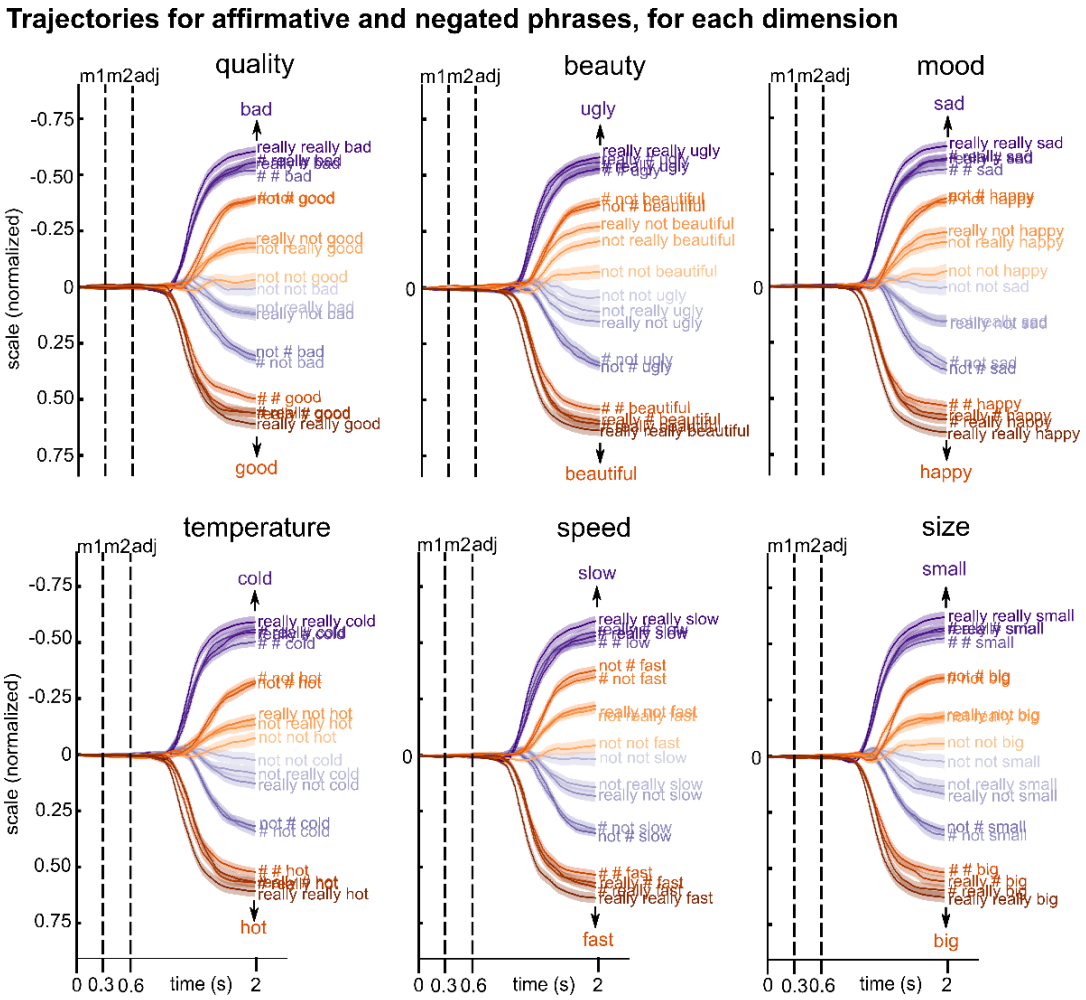


Figure S1. Trajectories for each dimension.

Behavioral trajectories for low (purples) and high (oranges) antonyms over time, for each 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.

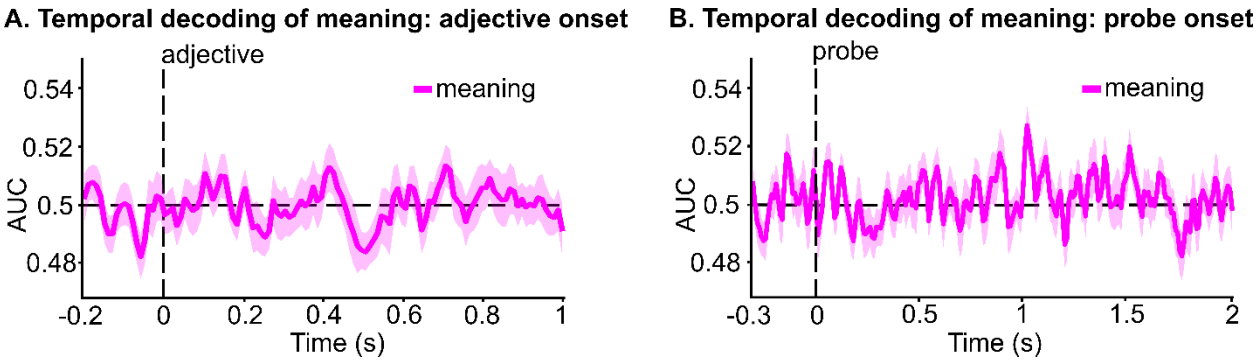


Figure S2. Temporal decoding of composed meaning.

We trained estimators on phrases where the 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 "goodness" dimension, classes were: [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". **A.** A temporal decoding analysis time-locked to the onset of the adjective did 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.** A temporal decoding analysis time-locked to the onset of the probe did 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 (A) and the probe (B); each line and shading represent participants' mean ± SEM.

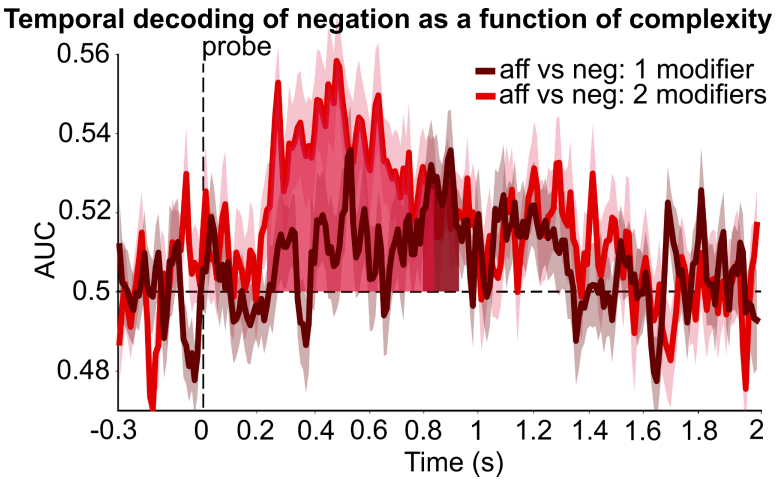


Figure S3. Temporal decoding of negation as a function of number of modifiers (i.e., complexity), time-locked to the onset of the probe.

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 again 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 negation 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.

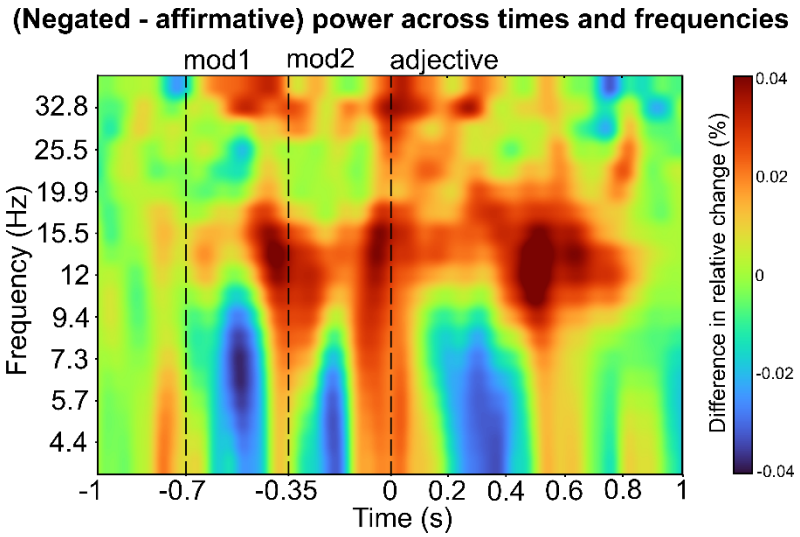


Figure S4. Differences between negated and affirmative phrases across time and frequencies.

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 (mod1).

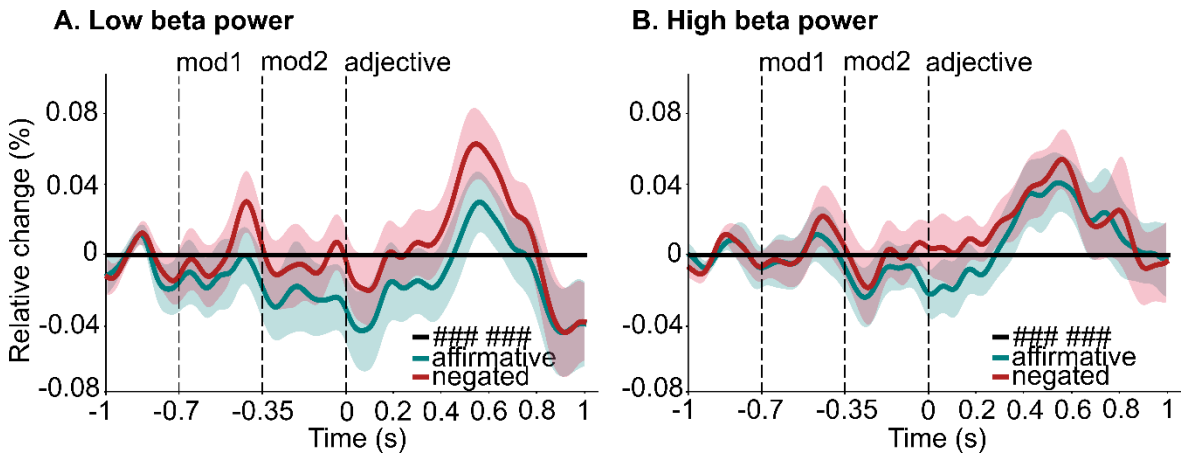


Figure S5. Low and high beta power for negated and affirmative phrases across time.

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 (20-30 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 word 1) 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.

References

Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y., & Ochiai, M. (2001). Reduction of non-periodic environmental magnetic noise in MEG measurement by Continuously Adjusted Least squares Method. IEEE Transactions on Applied Superconductivity, 11(1 I), 669–672. https://doi.org/10.1109/77.919433

Alemanno, F., Houdayer, E., Cursi, M., Velikova, S., Tettamanti, M., Comi, G., Cappa, S. F., & Leocani, L. (2012). Action-related semantic content and negation polarity modulate motor areas during sentence reading: An event-related desynchronization study. Brain Research, 1484, 39–49. https://doi.org/10.1016/j.brainres.2012.09.030

Andersen, L. M. (2018). 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(JAN). https://doi.org/10.3389/fnins.2018.00006

Aravena, P., Delevoye-Turrell, Y., Deprez, V., Cheylus, A., Paulignan, Y., Frak, V., & Nazir, T. (2012). Grip force reveals the context sensitivity of language-induced motor activity during "action words" processing: Evidence from sentential negation. PLoS ONE, 7(12), e50287. https://doi.org/10.1371/journal.pone.0050287

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39(3), 445–459. https://doi.org/10.3758/BF03193014

Bartoli, E., Tettamanti, A., Farronato, P., Caporizzo, A., Moro, A., Gatti, R., Perani, D., & Tettamanti, M. (2013). The disembodiment effect of negation: Negating action-related sentences attenuates their interference on congruent upper limb movements. Journal of Neurophysiology, 109(7), 1782–1792. https://doi.org/10.1152/jn.00894.2012

Bastiaansen, M. C. M., Linden, M. van der, Keurs, M., ter Dijkstra, T., & Hagoort, P. (2005). Theta responses are involved in lexical—Semantic retrieval during language processing. Journal of Cognitive Neuroscience, 17(3), 530–541. https://doi.org/10.1162/0898929053279469

Beltrán, D., Liu, B., & de Vega, M. (2021). Inhibitory mechanisms in the processing of negations: A neural reuse hypothesis. Journal of Psycholinguistic Research, 50(6), 1243–1260. https://doi.org/10.1007/s10936-021-09796-x

Beltrán, D., Morera, Y., García-Marco, E., & De Vega, M. (2019). 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(JULY), 1–14. https://doi.org/10.3389/fpsyg.2019.01782

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. Journal of Neuroscience, 31(8), 2801–2814. https://doi.org/10.1523/JNEUROSCI.5003-10.2011

Bemis, D. K., & Pylkkänen, L. (2013). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. Cerebral Cortex, 23(8), 1859–1873. https://doi.org/10.1093/cercor/bhs170

Berezutskaya, J., Freudenburg, Z. V., Ambrogioni, L., Güçlü, U., van Gerven, M. A. J., & Ramsey, N. F. (2020). Cortical network responses map onto data-driven features that capture visual semantics of movie fragments. Scientific Reports, 10(1), 12077. https://doi.org/10.1038/s41598-020-68853-y

Bianchi, I., Savardi, U., Burro, R., & Torquati, S. (2011). Negation and psychological dimensions. Journal of Cognitive Psychology, 23(3), 275–301. https://doi.org/10.1080/20445911.2011.493154

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex, 19(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. Psychological Review, 82(1), 45–73. https://doi.org/10.1037/h0076248

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling syntax and semantics in the brain with deep networks. http://arxiv.org/abs/2103.01620

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance, 9, 88–97. https://doi.org/10.1016/j.jbef.2015.12.001

Chen, Y., Shimotake, A., Matsumoto, R., Kunieda, T., Kikuchi, T., Miyamoto, S., Fukuyama, H., Takahashi, R., Ikeda, A., & Lambon Ralph, M. A. (2016). The “when” and “where” of

semantic coding in the anterior temporal lobe: Temporal representational similarity analysis of electrocorticogram data. Cortex, 79, 1–13. https://doi.org/10.1016/j.cortex.2016.02.015

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. Cognitive Psychology, 3(3), 472–517. https://doi.org/10.1016/0010-0285(72)90019-9

Colston, H. L. (1999). "Not good" is "bad," but "not bad" is not "good": An analysis of three accounts of negation asymmetry. Discourse Processes, 28(3), 237–256. https://doi.org/10.1080/01638539909545083

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron, 26(1), 55–67. https://doi.org/10.1016/S0896-6273(00)81138-1

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. Cognitive Science, 35(5), 983–996. https://doi.org/10.1111/j.1551-6709.2010.01164.x

Darley, E. J., Kent, C., & Kazanina, N. (2020). A 'no' with a trace of 'yes': A mouse-tracking study of negative sentence processing. Cognition, 198(December 2018), 104084. https://doi.org/10.1016/j.cognition.2019.104084

De Vega, M., Morera, Y., León, I., Beltrán, D., Casado, P., & Martín-Loeches, M. (2016). Sentential negation might share neurophysiological mechanisms with action inhibition. Evidence from frontal theta rhythm. Journal of Neuroscience, 36(22), 6002–6010. https://doi.org/10.1523/JNEUROSCI.3736-15.2016

Demos, R. (1917). Discussion of a certain type of negative proposition. Mind, 26(102), 188–196. https://doi.org/10.5840/philinquiry2016403/423

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. Nature Neuroscience, 19(1), 158–164. https://doi.org/10.1038/nn.4186

Djokic, V., Maillard, J., Bulat, L., & Shutova, E. (2019). Modeling affirmative and negated action processing in the brain with lexical and compositional semantic models. 5155–5165. https://doi.org/10.18653/v1/p19-1508

Dotan, D., & Dehaene, S. (2013). How do we convert a number into a finger trajectory? Cognition, 129(3), 512–529. https://doi.org/10.1016/j.cognition.2013.07.007

Dudschig, C., & Kaup, B. (2018). 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(5), 716–728. https://doi.org/10.1037/xhp0000481

Dudschig, C., Mackenzie, I. G., Leuthold, H., & Kaup, B. (2018). Environmental sound priming: Does negation modify N400 cross-modal priming effects? Psychonomic Bulletin and Review, 25(4), 1441–1448. https://doi.org/10.3758/s13423-017-1388-3

Ettinger, A. (2020). 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. https://doi.org/10.1162/tacl_a_00298

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. Proceedings of the National Academy of Sciences of the United States of America, 113(41), E6256–E6262. https://doi.org/10.1073/pnas.1612132113

Foroni, F., & Semin, G. R. (2013). Comprehension of action negation involves inhibitory simulation. Frontiers in Human Neuroscience, 7(MAY), 1–7. https://doi.org/10.3389/fnhum.2013.00209

Fraenkel, T., & Schul, Y. (2008). The meaning of negated adjectives. Intercultural Pragmatics, 5(4), 517–540. https://doi.org/10.1515/IPRG.2008.025

Fyshe, A., Sudre, G., Wehbe, L., Rafidi, N., & Mitchell, T. M. (2019). The lexical semantics of adjective–noun phrases in the human brain. Human Brain Mapping, 40(15), 4457–4469. https://doi.org/10.1002/hbm.24714

Gallese, V., & Lakoff, G. (2005). The brain’s concepts: The role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology, 22(3–4), 455–479. https://doi.org/10.1080/02643290442000310

Ghio, M., Haegert, K., Vaghi, M. M., & Tettamanti, M. (2018). 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(1752), 7–10. https://doi.org/10.1098/rstb.2017.0124

Giora, R. (2006). Anything negatives can do affirmatives can do just as well, except for some metaphors. Journal of Pragmatics, 38(7), 981–1014. https://doi.org/10.1016/j.pragma.2005.12.006

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience, 7(7 DEC), 1–13. https://doi.org/10.3389/fnins.2013.00267

Grice, H. P. (1975). Logic and Conversation. In P. Cole, & J. L. Morgan. (Eds.), Syntax and Semantics, Vol. 3, Speech Acts (pp. 41-58). New York: Academic Press.

Gwilliams, L., & King, J. R. (2020). Recurrent processes support a cascade of hierarchical decisions. ELife, 9, 1–20. https://doi.org/10.7554/ELIFE.56603

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. NeuroImage, 30(4), 1383–1400. https://doi.org/10.1016/j.neuroimage.2005.11.048

Horn, Laurence R. (1989). A natural history of negation. University of Chicago Press.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453–458. https://doi.org/10.1038/nature17637

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. Journal of Verbal Learning and Verbal Behavior, 10(3), 244–253. https://doi.org/10.1016/S0022-5371(71)80051-8

Kaup, B. (2001). Negation and its impact on the accessibility of text information. Memory and Cognition, 29(7), 960–967. https://doi.org/10.3758/BF03195758

Kaup, B., & Dudschig, C. (2020a). Understanding Negation. The Oxford Handbook of Negation, April, 634–655. https://doi.org/10.1093/oxfordhb/9780198830528.013.33

Kaup, B., & Dudschig, C. (2020b). Understanding Negation: Issues in the processing of negation. In V. Déprez & M. T. Espinal (Eds.), The Oxford Handbook of Negation (pp. 634–655). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198830528.013.33

Kaup, B., & Zwaan, R. A. (2003). Effects of negation and situational presence on the accessibility of text information. Journal of Experimental Psychology: Learning Memory and Cognition, 29(3), 439–446. https://doi.org/10.1037/0278-7393.29.3.439

Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A., & Ldtke, J. (2007). Experiential simulations of negated text information. Quarterly Journal of Experimental Psychology, 60(7), 976–990. https://doi.org/10.1080/17470210600823512

King, J. R., Pescetelli, N., & Dehaene, S. (2016). Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. Neuron, 92(5), 1122–1134. https://doi.org/10.1016/j.neuron.2016.10.051

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (n.d.). What' s new in Psychtoolbox-3?

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology, 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2016). The neural and computational bases of semantic cognition. Nature Reviews Neuroscience, 18(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Lau, E. F., Gramfort, A., Hämäläinen, M. S., & Kuperberg, G. R. (2013). Automatic semantic facilitation in anterior temporal cortex revealed through multimodal neuroimaging. Journal of Neuroscience, 33(43), 17174–17181. https://doi.org/10.1523/JNEUROSCI.1018-13.2013

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. Nature Reviews Neuroscience, 9(12), 920–933. https://doi.org/10.1038/nrn2532

Liuzza, M. T., Candidi, M., & Aglioti, S. M. (2011). Do not resonate with actions: Sentence polarity modulates cortico-spinal excitability during action-related sentence reading. PLoS ONE, 6(2), 38–41. https://doi.org/10.1371/journal.pone.0016855

Lüdtke, J., Friedrich, C. K., De Filippis, M., & Kaup, B. (2008). Event-related potential correlates of negation in a sentence-picture verification paradigm. Journal of Cognitive Neuroscience, 20(8), 1355–1370. https://doi.org/10.1162/jocn.2008.20093

Luo, Y., Zhang, Y., Feng, X., & Zhou, X. (2010). Electroencephalogram oscillations differentiate semantic and prosodic processes during sentence reading. Neuroscience, 169(2), 654–664. https://doi.org/10.1016/j.neuroscience.2010.05.032

Lyons, J. (1995). Linguistic semantics: An introduction. New York, NY: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511810213

MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(4), 633–642. https://doi.org/10.1037/0278-7393.15.4.633

Maldonado, M., Dunbar, E., & Chemla, E. (2019). Mouse tracking as a window into decision making. Behavior Research Methods, 51(3), 1085–1101. https://doi.org/10.3758/s13428-018-01194-x

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods, 164(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martin, A. E., & Doumas, L. A. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. PLoS Biology, 15(3), 1–23. https://doi.org/10.1371/journal.pbio.2000663

Matchin, W., & Hickok, G. (2020). The Cortical Organization of Syntax. Cerebral Cortex, 30(3), 1481–1498. https://doi.org/10.1093/cercor/bhz180

Mayo, R., Schul, Y., & Burnstein, E. (2004). "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(4), 433–449. https://doi.org/10.1016/j.jesp.2003.07.008

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. Psychological Science, 19(12), 1213–1218. https://doi.org/10.1111/j.1467-9280.2008.02226.x

Noonan, K. A., Jefferies, E., Visser, M., & Lambon Ralph, M. A. (2013). Going beyond inferior prefrontal involvement in semantic control: Evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. Journal of Cognitive Neuroscience, 25(11), 1824–1850. https://doi.org/10.1162/jocn_a_00442

Oseki, Y., & Marantz, A. (2020). Modeling morphological processing in human magnetoencephalography. Proceedings of the Society for Computation in Linguistics, 3.

Palaz, B., Rhodes, R., & Hestvik, A. (2020). Informative use of "not" is N400-blind. Psychophysiology, 57(12). https://doi.org/10.1111/psyp.13676

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. Proceedings of the National Academy of Sciences, 108(6), 2522–2527. https://doi.org/10.1073/pnas.1018711108

Papeo, L., & de Vega, M. (2020). The neurobiology of lexical and sentential negation. The Oxford Handbook of Negation, 739–756. https://doi.org/10.1093/oxfordhb/9780198830528.013.44

Papeo, L., Hochmann, J.-R., & Battelli, L. (2016). The Default Computation of Negated Meanings. Journal of Cognitive Neuroscience, 28(12), 1980–1986. https://doi.org/10.1162/jocn_a_01016

Parrish, A., & Pylkkänen, L. (2022). Conceptual combination in the LATL with and without syntactic composition. Neurobiology of Language, 3(1), 46–66. https://doi.org/10.1162/nol_a_00048

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience, 8(12), 976–987. https://doi.org/10.1038/nrn2277

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology, 70, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pinheiro-Chagas, P., Dotan, D., Piazza, M., & Dehaene, S. (2017). Finger tracking reveals the covert stages of mental arithmetic. Open Mind, 1(1), 30–41. https://doi.org/10.1162/opmi_a_00003

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. Nature Neuroscience, 24(11), 1628–1636. https://doi.org/10.1038/s41593-021-00921-6

Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. Trends in Cognitive Sciences, 17(9), 458–470. https://doi.org/10.1016/j.tics.2013.06.004

Pulvermüller, F., Assadollahi, R., & Elbert, T. (2001). Neuromagnetic evidence for early semantic access in word recognition. European Journal of Neuroscience, 13(1), 201–205. https://doi.org/10.1046/j.0953-816X.2000.01380.x

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. Brain and Language, 110(2), 81–94. https://doi.org/10.1016/j.bandl.2008.12.001

Pylkkänen, L., & Marantz, A. (2003). Tracking the time course of word recognition with MEG. Trends in Cognitive Sciences, 7(5), 187–189. https://doi.org/10.1016/S1364-6613(03)00092-5

Russell, B. (1948). Human knowledge: Its scope and limits. New York: Simon & Schuster.

Schaller, F., Weiss, S., & Müller, H. M. (2017). EEG beta-power changes reflect motor involvement in abstract action language processing. Brain and Language, 168, 95–105. https://doi.org/10.1016/j.bandl.2017.01.010

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. Behavior Research Methods, 46(1), 95–111. https://doi.org/10.3758/s13428-013-0345-y

Speranza, J. L., & Horn, L. R. (2010). A brief history of negation. Journal of Applied Logic, 8(3), 277–301. https://doi.org/10.1016/j.jal.2010.04.001

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. NeuroImage, 62(1), 451–463. https://doi.org/10.1016/j.neuroimage.2012.04.048

Supp, G. G., Schlögl, A., Gunter, T. C., Bernard, M., Pfurtscheller, G., & Petsche, H. (2004). Lexical memory search during N400: Cortical couplings in auditory comprehension: NeuroReport, 15(7), 1209–1213. https://doi.org/10.1097/00001756-200405190-00026

Teige, C., Mollo, G., Millman, R., Savill, N., Smallwood, J., Cornelissen, P. L., & Jefferies, E. (2018). Dynamic semantic cognition: Characterising coherent and controlled conceptual retrieval through time using magnetoencephalography and chronometric transcranial magnetic stimulation. Cortex, 103, 329–349. https://doi.org/10.1016/j.cortex.2018.03.024

Tettamanti, M., Manenti, R., Della Rosa, P. A., Falini, A., Perani, D., Cappa, S. F., & Moro, A. (2008). Negation in the brain: Modulating action representations. NeuroImage, 43(2), 358–367. https://doi.org/10.1016/j.neuroimage.2008.08.004

Tomasino, B., Weiss, P. H., & Fink, G. R. (2010). To move or not to move: Imperatives modulate action-related verb processing in the motor system. Neuroscience, 169(1), 246–258. https://doi.org/10.1016/j.neuroscience.2010.04.039

Vandenberghe, R., Price, C., Wise, R., Josephs, O., & Frackowiak, R. S. J. (1996). Functional anatomy of a common semantic system for words and pictures. Nature, 383(6597), 254–256. https://doi.org/10.1038/383254a0

Wagner, J., Wessel, J. R., Ghahremani, A., & Aron, A. R. (2018). 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(1), 107–118. https://doi.org/10.1162/jocn_a_01183

Weiss, S., & Mueller, H. M. (2012). "Too many betas do not spoil the broth": The role of beta brain oscillations in language processing. Frontiers in Psychology, 3(JUN), 1–15. https://doi.org/10.3389/fpsyg.2012.00201

Weiss, S., & Rappelsberger, P. (1996). 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(1), 17–20. https://doi.org/10.1016/0304-3940(96)12581-7

Xiang, M., Grove, J., & Giannakidou, A. (2016). Semantic and pragmatic processes in the comprehension of negation: An event related potential study of negative polarity sensitivity. Journal of Neurolinguistics, 38, 71–88. https://doi.org/10.1016/j.jneuroling.2015.11.001

Zhang, L., & Pylkkänen, L. (2018). Semantic composition of sentences word by word: MEG evidence for shared processing of conceptual and logical elements. Neuropsychologia, 119(April), 392–404. https://doi.org/10.1016/j.neuropsychologia.2018.08.016